

PUBLICATIONS DE L'UNIVERSITÉ DE NANCY

COLLECTION DES THÈSES DE LA FACULTÉ DES SCIENCES

MATHÉMATIQUES

PREMIÈRE THÈSE

**Suritération des grands systèmes linéaires
quelques applications aux équations
aux dérivées partielles**

DEUXIÈME THÈSE

Le traitement des listes en PL 1

**Présentées à la Faculté des Sciences de l'Université de Nancy
pour obtenir le grade de Docteur ès-Sciences Mathématiques**

par Philippe NOËL

soutenues le 24 Février 1967

devant la Commission d'Examen

JURY :

M. LEGRAS, Président

M. EYMARD

M. DEPAIX

M. PAIR

} Examinateurs

1967



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41555012_0042

PUBLICATIONS DE L'UNIVERSITÉ DE NANCY

COLLECTION DES THÈSES DE LA FACULTÉ DES SCIENCES

MATHÉMATIQUES

PREMIÈRE THÈSE

Suritération des grands systèmes linéaires quelques applications aux équations aux dérivées partielles

DEUXIÈME THÈSE

Le traitement des listes en PL 1

Présentées à la Faculté des Sciences de l'Université de Nancy
pour obtenir le grade de Docteur ès-Sciences Mathématiques

par **Philippe NOËL**

soutenues le 24 Février 1967

devant la Commission d'Examen

JURY :

M. LEGRAS, Président

M. EYMARD

M. DEPAIX

M. PAIR

} **Examineurs**

1967

Doyen : M. AUBRY

Assesseur : M. GAY

Doyens honoraires : MM. CORNUBERT - DELSARTE - URION - ROUBAULT.

Professeurs honoraires : MM. CROZE - RAYBAUD - LAFFITIE - LERAY - JULY - LAPORTE - EICHHORN - GODEMENT - DUBREUIL - L. SCHWARTZ - DIEUDONNE - DE MALLEMANN - LONGGHAMBON - LETORT - DODE - GAUTHIER - GOUDET - OLMER - CORNUBERT - CHAPELLE - GUERIN - WAHL.

Maîtres de conférences honoraires : MM. LIENHART - PIERRET.

Professeurs

MM. URION	Chimie biologique	HEROLD	Chimie minér, indus.
DELSARTE	Analyse supérieure	SCHWARTZ B.	Exploitation minière
ROUBAULT	Géologie	MANGENOT	Botanique
CHAPELLE	Méca. Rationnelle	GAYET	Physiologie
VEILLET	Biologie animale	BONVALET	Méca. appliquée
BARRIOL	Chimie théorique	HADNI	Physique
BIZETTE	Physique	* BASTICK	Chimie
GUILLIEN	Electronique	DUCHAUFOR	Pédologie
GILBERT	Chimie physique	GARNIER	Agronomie
LEGRAS	Méca. Rationnelle	NEEL	Chimie org. indus.
BOLFA	Minér. & crist.	BERNARD	Géologie appliquée
NICLAUSE	Chimie	* CHAMPIER	Physique
FAIVRE	Phys. appliquée	* GAY	Chimie biologique
AUBRY	Chimie minérale	STEPHAN	Zoologie
COPPENS	Radiogéologie	* CONDE	Zoologie
* MALAPRADE	Chimie	* WERNER	Botanique
DUVAL	Chimie	EYMARD	Calcul dif. & int.
FRUHLING	Physique	LEVISALLES	Chimie organique
HILLY	Géologie	Mme HERVE	M.M.P.
LE GOFF	Génie chimique	* GOSSE	Méca. physique
SUHNER	Phys. expériment.	FELDEN	Physique
CHAPON	Chimie biologique	* DAVOINE	E.N.S.M.I.M.

* Professeur titulaire à titre personnel.

Maîtres de Conférences

Mme BASTICK	Chimie M.P.C.	JACQUIN	Pédol. & chim. agr.
MM. GUDEFIN	Physique	MAINARD	Phys. M.G.P.
ROCCI	Géologie	CACHAN	Ento. appl. E.N.S.A.
VUILLAUME	Psychophysio.	MARTIN	Chimie M.P.C.
FRENTZ	Biol. animale	PAULMIER	Méca. expérimentale
HORN	Phys. propédeutique	PROTAS	Minéralogie
LAFON	Phys. (I.S.I.N.)	JOZEFOWICZ	Physico-chimie
MARI	Chimie (I.S.I.N.)	JURAIN	Géologie S.P.C.N.
AUROUZE	Géologie	N...	Proba. & statistiques
DEVIOT	Physi du solide	N...	Mécanique I.S.I.N.
FLECHON	Physique M.P.C.	N...	Mathématiques
Mle HUET *	Math. S.P.C.N.	N...	Génie chimique
MM. VIGNES	Métallurgie	N...	Mathématiques M.P.C.
BALESDENT	Thermodynam. chim.	N...	Mathématiques S.P.C.N.
BLAZY	Minéral. appl. E.N.S.G.	N...	Mathématiques appliquées
JANOT	Phys. M.P.C. (Epinal)	N...	Méca. des fluides
		N...	Physiol. animale

* en disponibilité

Secrétaire principal : C. CARON

A ma Femme,

Ce travail a été réalisé sous la direction de Monsieur le Professeur J. LEGRAS, Directeur de l'Institut de Calcul Automatique de l'Université de NANCY.

Je tiens à lui exprimer toute ma reconnaissance pour la bienveillante attention et le soutien constant qu'il n'a cessé de m'accorder.

La participation au Jury de Monsieur P. EYMARD, Professeur à la Faculté des Sciences de NANCY, est un grand honneur pour moi. Je l'assure de ma respectueuse reconnaissance.

Monsieur Cl. PAIR, chargé d'enseignement à la Faculté des Sciences de NANCY, a bien voulu me proposer le second sujet. Je le remercie très sincèrement pour sa sollicitude à mon égard.

Je remercie vivement Monsieur M. DEPAIX, chargé d'enseignement à la Faculté des Sciences de NANCY pour sa participation au Jury.

Mes remerciements vont enfin à tous mes camarades pour leur sympathique collaboration.

Ce travail a été effectué dans le cadre du contrat n° 38764, passé avec la Direction des Recherches et Moyens d'Essais.

INTRODUCTION

Les méthodes numériques utilisées pour résoudre un système linéaire $[A] [X] = [B]$ sont classiquement divisées en méthodes directes et méthodes itératives.

Les premières (référence [1]) donnent directement, par transformation de la matrice du système, la solution en un nombre fini d'opérations arithmétiques, ce nombre dépendant de la dimension du système. L'importance de l'influence des erreurs de troncature rend dangereuse l'utilisation de telles méthodes, dès que le système atteint des grandeurs non négligeables.

Les méthodes itératives consistent en la répétition d'un algorithme plus ou moins simple fournissant la solution $[X]$, en tant que limite d'une suite vectorielle de terme $[X_k]$.

Si l'itéré $[X_k]$ est fonction de $[A]$, $[B]$, $[X_{k-1}]$, ... $[X_{k-r}]$, l'itération est dite de degré r . Ce type d'itération est rarement utilisé, car il nécessite la mise en réserve des vecteurs $[X_{k-1}]$, $[X_{k-2}]$... et $[X_{k-r}]$. Si des itérations de degré 2 sont parfois utilisées, les plus communes sont de degré 1.

$$X_k = F_k ([A], [B], [X_{k-1}]).$$

Le procédé itératif est linéaire si F_k est fonction linéaire de $[X_{k-1}]$. Il est dit stationnaire si F_k est indépendant de k .

La fonction linéaire de degré 1 la plus générale est :

$$[X_k] = [H_k] [X_{k-1}] + [M_k] [B]$$

les matrices $[M_k]$ et $[H_k]$, fonctions de $[A]$ et $[B]$ étant liées par la relation $[H_k] + [M_k] [A] = [I]$.

Parmi les méthodes itératives les plus classiques, nous pouvons citer la méthode des déplacements simultanés (ou méthode de Jacobi), la méthode des déplacements successifs, et la méthode du gradient. Les descriptions détaillées concernant ces méthodes peuvent être trouvées en référence [1], [2] ou [3].

Les méthodes du gradient utilisent des itérations linéaires, soit stationnaires, soit non stationnaires.

$$[X_k] = ([I] - \alpha [A]) [X_{k-1}] + \alpha [B]$$

ou

$$[X_k] = ([I] - \alpha_k [A]) [X_{k-1}] + \alpha_k [B]$$

Les itérations proposées par L.F. Richardson en 1910 sont du type ci-dessus. La méthode, qui porte son nom, a connu un rapide essor dès l'avènement des calculatrices électroniques, vers 1950. D. Young (référence [4]) en a été l'un des plus fervents utilisateurs. Les difficultés rencontrées dans son emploi, difficultés dues à l'influence des erreurs de troncature,

rendaient ce procédé souvent instable sur le plan numérique. E. Stiefel (référence [5]), en transformant les itérations de Richardson en des itérations linéaires du second degré, a éliminé ce phénomène d'instabilité. Mais ces itérations ont, vis à vis de celles de Richardson, l'inconvénient de nécessiter une plus grande occupation de mémoires.

Une étude expérimentale de la méthode de Richardson nous a permis de réduire considérablement ce phénomène d'instabilité qui la caractérisait, et de rendre cette méthode opérationnelle et compétitive.

Après avoir décrit dans la première partie la méthode de Richardson et la technique d'emploi que nous préconisons, nous présentons dans la deuxième partie un certain nombre d'applications qui montrent l'intérêt de cette méthode.

PREMIERE PARTIE

METHODES ET TECHNIQUES

I - 1. METHODES

1. 1. 1. Position du problème.

Nous nous proposons de résoudre un système linéaire de n équations à n inconnues

$$[A] [X] = [B] \quad (I - 1)$$

en résolvant par approximations successives le nouveau système

$$[X] = [P_p(A)] [X] - [Q_{p-1}(A)] [B] \quad (II - 1)$$

où $P_p(A)$ et $Q_{p-1}(A)$ sont deux polynômes matriciels.

Le polynôme $P_p(A)$ sera choisi de façon à rendre convergente la suite

$$[X_k] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B] \quad (III - 1)$$

le rôle de $[Q_{p-1}(A)]$ étant de rendre la limite de cette suite identique à la solution de (I-I) c'est-à-dire $[A]^{-1} [B]$.

Nous nous sommes limités au cas où la matrice $[A]$ est diagonalisable et à éléments réels, et où ses valeurs propres sont toutes réelles et de même signe.

1. 1. 2. Notations utilisées.

Les $\lambda_i, i=1, \dots, n$ représenteront dans toute cette étude les valeurs propres de $[A]$.

$[A]$ étant par hypothèse diagonalisable, nous ferons souvent appel à la relation :

$$[A] = [\Omega] \begin{bmatrix} \dots & \lambda_i & \dots \end{bmatrix} [\Omega]^{-1}$$

où $\begin{bmatrix} \dots & \lambda_i & \dots \end{bmatrix}$ représente la matrice diagonale dont les λ_i sont les éléments de la diagonale. [Ces éléments ne sont pas nuls, $[A]$ étant inversible ;] et où $[\Omega]$ est la matrice carrée formée à partir des vecteurs propres de $[A]$, la $i^{\text{ème}}$ colonne étant le vecteur propre normé associé à la $i^{\text{ème}}$ valeur propre λ_i .

Si $p_p(x) = \alpha_0 + \alpha_1 x + \dots + \alpha_p x^p = \sum_{i=0}^p \alpha_i x^i$, le polynôme matriciel associé $P_p(A)$ est par définition :

$$[P_p(A)] = \alpha_0 [I] + \alpha_1 [A] + \dots + \alpha_p [A]^p = \sum_{i=0}^p \alpha_i [A]^i$$

$[P_p(A)]$ peut se mettre sous la forme :

$$[P_p(A)] = [\Omega] \begin{bmatrix} \dots & p_p(\lambda_i) & \dots \end{bmatrix} [\Omega]^{-1}$$

car $[P_p(A)]$ admet pour vecteurs propres les vecteurs propres de $[A]$ associés aux valeurs propres $p_p(\lambda_i)$ ($i=1, \dots, n$).

Nous adopterons toujours comme norme vectorielle la norme euclidienne

$$\|X\| = \sqrt{\sum x_i^2}$$

La norme matricielle utilisée sera la norme qui lui est associée

$$\|A\| = \max_i \sqrt{|\lambda_i^*|} \text{ avec } \lambda_i^* = \text{valeurs propres de } [A^t][A].$$

Si la matrice de transformation $[\Omega]$ est orthogonale, on vérifie que

$$\|A\| = \max_i |\lambda_i| = \rho(A).$$

$\rho(A)$ est le rayon spectral de $[A]$ et on a toujours $\rho(A) \leq \|A\|$.

1. 1. 3. Détermination des polynômes matriciels $P_p(A)$ et $Q_{p-1}(A)$.

1. 1. 3. 1. Convergence de la suite $[X_k] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B]$

L'étude de la convergence de la suite $[X_k]$ peut se faire en étudiant la série de terme général $[U_k]$ définie par :

$$[U_k] = [X_k] - [X_{k-1}]$$

car, à la convergence de la suite $[X_k]$ est associée celle de la série $[U_k]$,

$$[X_k] \text{ valant } [U_1] + [U_2] + \dots + [U_k] + [X_0]$$

$$[U_k] = [X_k] - [X_{k-1}] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B] - [P_p(A)] [X_{k-2}] + [Q_{p-1}(A)] [B]$$

$$\implies [U_k] = [P_p(A)] [U_{k-1}] \quad (\text{IV} - 1)$$

$$\text{De plus : } [U_{k-1}] = [P_p(A)] [U_{k-2}]$$

$$\implies [U_k] = [P_p(A)]^2 [U_{k-2}]$$

et de proche en proche :

$$[U_k] = [P_p(A)]^{k-1} [U_1] \quad (\text{V} - 1)$$

La série $[U_k]$ converge si, et seulement si, $\|U_k\|$ converge ; mais

$$\|U_k\| = \|[P_p(A)] [U_{k-1}]\| \leq \|P_p(A)\| \|U_{k-1}\|.$$

On retrouve ainsi une condition suffisante de convergence

$$\|P_p(A)\| < 1$$

La condition nécessaire et suffisante est donnée par :

$$\rho(P_p(A)) < 1 \quad (\text{voir réf. [21]})$$

$$\implies \max_i |p_p(\lambda_i)| < 1$$

1. 1. 3. 2. Equivalence des systèmes (I-1) et (II-1)

Si les conditions de convergence sont réalisées, c'est-à-dire si $\rho(P_p(A)) = \max_i |p_p(\lambda_i)| < 1$, la suite $[X_k] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B]$ converge vers le vecteur $[X]$ solution de $[X] = [P_p(A)] [X] - [Q_{p-1}(A)] [B]$.

a) - Si cette solution est aussi solution de $[A] [X] = [B]$, (II-1) devient $[X] = [P_p(A)] [X] - [Q_{p-1}(A)] [A] [X]$

$$\iff [[Q_{p-1}(A)] [A] - [P_p(A) - I]] [X] = [0].$$

$[X]$ étant unique et non identiquement nul, cette condition sera réalisée, quel que soit $[P_p(A)]$ si

$$[Q_{p-1}(A)] = [[P_p(A)] - [I]] [A]^{-1} \quad (VI-1)$$

$$\text{Or, } [P_p(A)] = \sum_{i=0}^p \alpha_i [A]^i$$

$$\implies [Q_{p-1}(A)] = [\alpha_0 - 1] [A]^{-1} + \sum_{i=1}^p \alpha_i [A]^{i-1}$$

$[Q_{p-1}(A)]$ sera un polynôme matriciel si $\alpha_0 = 1$, c'est-à-dire si $p_p(x)$ vérifie $p_p(0) = 1$; d'où les relations devant lier $p_p(x)$ et $q_{p-1}(x)$, polynômes auxquelles sont associés $P_p(A)$ et $Q_{p-1}(A)$.

$$p_p(0) = 1$$

$$q_{p-1}(x) = \frac{p_p(x) - 1}{x}$$

b) - **Réciproquement**, montrons que, si $[Q_{p-1}(A)] = [P_p(A) - I] [A]^{-1}$ la solution de (III-1) est $[A^{-1}] [B]$, lorsqu'il y a convergence.

$$(II-1) \implies [X] = [P_p(A)] [X] - [[P_p(A)] - [I]] [A^{-1}] [B]$$

$$\implies [[P_p(A)] - [I]] [X] = [[P_p(A)] - [I]] [A^{-1}] [B] \quad (VII-1)$$

La matrice $[P_p(A)] - [I]$ est diagonalisable $[A]$ l'étant, et ses valeurs propres ne sont pas nulles car $[P_p(A)]$ a été choisi avec $\rho(P_p(A)) < 1$.

Cette matrice $[P_p(A)] - [I]$ est inversible. En prémultipliant les deux membres de l'égalité (VII-1) par $[P_p(A) - I]^{-1}$ on retrouve bien $[X] = [A^{-1}] [B]$.

1. 1. 3. 4. Théorème fondamental

Les conditions de convergence de la suite (III-1) et d'équivalence des deux systèmes (I-1) et (II-1) peuvent être regroupées dans le théorème suivant :

Théorème : Si $p_p(x)$ et $q_{p-1}(x)$ sont deux polynômes liés par la relation $q_{p-1}(x) = (p_p(x) - 1) / x$, avec $p_p(0) = 1$, la suite matricielle de terme général

$$[X_k] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B]$$

converge vers la solution du système $[A] [X] = [B]$ si, et seulement si, $\rho(P_p(A)) = \max_i |p_p(\lambda_i)| < 1$.

1. 1. 4 Estimation de l'erreur commise.

Supposons la suite $[X_k]$ convergente vers $[X] = [A]^{-1} [B]$. Nous nous proposons d'étudier la précision obtenue à la $k^{\text{ème}}$ itération, en fonction de k et du polynôme choisi, à l'aide du vecteur

$$[E_k] = [X] - [X_k].$$

A cet effet, nous associons la série vectorielle déjà définie $[U_k] = [X_k] - [X_{k-1}]$ qui vérifie les deux relations

$$[U_k] = [P_p(A)] [U_{k-1}]$$

et $[U_k] = [P_p(A)]^{k-1} [U_1]$

Si la suite converge

$$[E_k] = [X] - [X_k] = \lim_{q \rightarrow \infty} \sum_{i=k+1}^q [[X_i] - [X_{i-1}]]$$

soit $[E_k] = \lim_{q \rightarrow \infty} \sum_{i=k+1}^q [U_i]$

$$\left[\begin{array}{l} [U_{k+1}] = [P_p(A)] [U_k] \\ [U_{k+2}] = [P_p(A)] [U_{k+1}] = [P_p(A)]^2 [U_k] \\ \dots\dots\dots \\ [U_{k+q}] = [P_p(A)] [U_{k+q-1}] = \dots = [P_p(A)]^q [U_k] \\ \sum_{i=k+1}^q [U_i] = [[P_p(A)] + [P_p(A)]^2 + \dots + [P_p(A)]^q] [U_k] \end{array} \right.$$

La matrice $[\eta_q] = [[P_p(A)] + \dots + [P_p(A)]^q]$, somme d'une série matricielle "géométrique" de "raison" $[P_p(A)]$ est définie, car $\rho([P_p(A)]) < 1$.

$$[E_k] = \lim_{q \rightarrow \infty} [\eta_q] [U_k] \text{ a donc un sens.}$$

D'où une estimation de la norme du vecteur $[E_k]$:

$$\|E_k\| < \lim_{q \rightarrow \infty} \|\eta_q\| \|U_k\|$$

Or, $[\eta_q] = [P_p(A) + \dots + P_p(A)^q]$

$$= [\Omega] \left[\dots p_p(\lambda_1) + p_p^2(\lambda_1) + \dots + p_p^q(\lambda_1) \dots \right] [\Omega]^{-1}$$

$$= [\Omega] \left[\dots p_p(\lambda_1) \left(\frac{1 - p_p^q(\lambda_1)}{1 - p_p(\lambda_1)} \right) \dots \right] [\Omega]^{-1}$$

Par hypothèse, $|p_p(\lambda_i)| < 1 \quad \forall i = 1, \dots n.$

$$\Rightarrow \lim_{q \rightarrow \infty} [\eta] = [\Omega] \left[\dots \frac{p(\lambda_i)}{1-p(\lambda_i)} \dots \right] [\Omega]^{-1}$$

$$\text{donc } \left\| \lim_{q \rightarrow \infty} [\eta_q] \right\| \leq \frac{h}{1-h} \cdot K \text{ avec } \begin{cases} h = \rho(P_p(A)) = \max |p_p(\lambda_i)| \\ K = \left\| \Omega \right\|, \left\| \Omega^{-1} \right\| \geq 1 \\ (K = 1 \text{ si } [\Omega] \text{ est orthogonale}) \end{cases}$$

D'où une première estimation de l'erreur à la $k^{\text{ème}}$ itération, en fonction de $[U_k] = [X_k] - [X_{k-1}]$, vecteur différence des deux derniers itérés calculés.

$$\boxed{\|E_k\| \leq \frac{h}{1-h} \cdot K \cdot \|U_k\|} \quad (\text{VIII} - 1)$$

$$\text{de plus } [U_k] = [P_p(A)]^{k-1} [U_1]$$

$$\longrightarrow \|U_k\| \leq h^{k-1} \|U_1\| \cdot K$$

En reportant dans (VIII - 1), nous obtenons la deuxième formule fondamentale, permettant de connaître théoriquement la précision obtenue (au sens de la norme de E_k), en fonction de U_1 .

$$\boxed{\|E_k\| \leq \frac{h^k}{1-h} \|U_1\| \cdot K^2} \quad (\text{IX} - 1)$$

Remarque :

1°) Si $[X_0] = \begin{bmatrix} \vdots \\ 0 \end{bmatrix}$ vecteur initial le plus souvent utilisé, la formule (IX - 1) devient :

$$\|E_k\| < \frac{h^k}{1-h} \|X_1\| \cdot K^2 \quad (\text{IXbis} - 1)$$

2°) Cette dernière relation (IX-1) permet théoriquement de donner, à partir de la connaissance de U_1 et de h , une estimation du nombre d'itérations k exigé pour atteindre la précision voulue sur $\|E_k\|$

$$k = \text{entier} \left(\frac{\text{Log } \|E_k\| - 2 \text{Log } K \|U_1\| + \text{Log } (1-h)}{\text{Log } (|h|)} \right) + 1$$

3°) Une autre relation peut aussi être intéressante pour préciser l'erreur au sens relatif, l'erreur au sens absolu étant donnée par (VIII - 1). En effet :

$$\begin{aligned} [E_k] &= [P_p(A)] [E_{k-1}] = [P_p(A)]^2 [E_{k-2}] \\ &= \dots = [P_p(A)]^k [E_0] \end{aligned}$$

$$\Rightarrow \boxed{\frac{\|X - X_k\|}{\|X - X_0\|} \leq h^k \cdot K}$$

4°) Toutes ces formules ne seront effectivement utilisables que lorsque $K=1$, c'est-à-dire $\|\Omega\| = \|\Omega^{-1}\| = 1$, $[A]$ définie positive par exemple.

1. 1. 5. Choix du polynôme de convergence $P_p(A)$.

Nous recherchons parmi tous les polynômes de degré P vérifiant les conditions du théorème fondamental celui qui assure la convergence la plus rapide, c'est-à-dire celui qui nécessite le moins d'itérations pour atteindre la précision voulue.

Les relations (VIII - 1) et (IX - 1) montrent que cette rapidité de convergence est fonction de $h = \max_i |p_p(\lambda_i)|$.

Deux types de problèmes peuvent être envisagés :

- celui pour lequel la matrice $[A]$ a toutes ses valeurs propres connues ;
- celui pour lequel ces valeurs, a priori, ne le sont pas.

Le premier cas est exceptionnel, et nous ne ferons que l'effleurer.

1. 1. 5. 1. Résolution d'un système linéaire pour lequel toutes les valeurs propres de la matrice sont connues.

Si les λ_i sont tous connus, le polynôme matriciel $P_p(A)$ choisi pourra être le polynôme associé à :

$$p_p(x) = \left(1 - \frac{x}{\lambda_1}\right) \left(1 - \frac{x}{\lambda_2}\right) \dots \left(1 - \frac{x}{\lambda_n}\right)$$

Ce polynôme de degré $p=n$ vaut bien 1 à l'origine, et ses racines sont égales aux valeurs propres.

$$\implies h = \max_i |p_p(\lambda_i)| = 0$$

Une itération doit donc théoriquement suffire à assurer la convergence et le vecteur obtenu $[X_1]$ doit être, aux erreurs de troncature près, strictement égal à $[A^{-1}] [B]$.

1. 1. 5. 2. Cas général - Les valeurs propres de la matrice $[A]$ ne sont pas toutes connues.

Dans la majorité des systèmes qui se présentent, les valeurs propres ne sont pas toutes connues. Nous supposons néanmoins que nous sommes en mesure de déterminer deux nombres α et β , respectivement majorant et minorant des valeurs propres.

$$\implies 0 < \beta \leq \lambda_{\min} \dots < \lambda_i < \dots \leq \lambda_{\max} \leq \alpha$$

(Rappelons que nous nous plaçons dans le cas où toutes les valeurs propres sont positives).

Si $p_p(x)$ est le polynôme associé à $P_p(A)$, nous savons que :

$$h = \max_i |p_p(\lambda_i)| \leq \max_{x \in [\beta, \alpha]} |p_p(x)|$$

Le polynôme que nous choisissons sera, parmi tous les polynômes de même degré p et valant 1 pour $x=0$, celui qui est tel que la maximum qu'il prend sur $[\beta, \alpha]$ en valeur absolue est minimum. On sait qu'un tel polynôme s'obtient à partir des polynômes de Tchebycheff, dont nous allons rappeler les principales propriétés.

1. 1. 5. 3. Polynômes de Tchebycheff de première espèce.

Définition :

Le polynôme $T_p(x)$, de degré p , se définit par l'égalité :

$$T_p(x) = \begin{cases} \cos p(\text{Arcos } x) & \text{si } x \in [-1, +1] \\ \varepsilon^p \text{ch } p(\text{Arg}^t \text{ch } |x|) & \text{si } |x| \geq 1. \end{cases}$$

$$\varepsilon = \begin{cases} -1 & \text{si } x \leq -1 \\ +1 & \text{si } x \geq +1 \end{cases}$$

Propriétés essentielles

1°) Variations de $T_p(x)$ pour $x \in]-\infty, +\infty[$

- pour $x \in [-1, +1]$, la courbe représentative oscille $p-1$ fois sur cet intervalle entre -1 et $+1$, et s'annule p fois aux points d'abscisse

$$x_k = \cos \frac{(2k-1)\pi}{2p} \quad k = 1, 2, \dots, p$$

Les valeurs $+1$ et -1 étant atteintes aux points d'abscisse, $x_i = \cos \frac{i\pi}{p}$.

- dès que x sort de cet intervalle, $T_p(x)$ tend très rapidement vers

$$\begin{cases} +\infty & \text{pour } x > 1 \\ (-1)^p \infty & \text{pour } x < -1. \end{cases}$$

2°) Racines de l'équation $T_p(\omega) = C$

C étant une constante numérique.

a) si $|C| < 1 \implies p$ racines réelles données par :

$$\omega_k = \omega_0 \cos \frac{2k\pi}{p} - \sqrt{1 - \omega_0^2} \sin \frac{2k\pi}{p}$$

avec $\omega_0 = \cos \left(\frac{\gamma}{p} \right)$ si $\gamma = \text{Arcos } C$.

b) si $|C| > 1$:

1 ou 2 racines réelles suivant la parité de p et le signe de C.

$$\text{Posons } V(C, p) = \frac{1}{2} \left[(C + \sqrt{C^2 - 1})^{1/p} + (C - \sqrt{C^2 - 1})^{1/p} \right]$$

$$\text{si } C > 1 \text{ et p pair : 2 racines } \begin{cases} \omega_1 = V(C, p) \\ \omega_2 = -\omega_1 \end{cases}$$

$$p \text{ impair : 1 racine : } \omega_1 = V(C, p)$$

si $C < -1$ et p pair : pas de racine

$$p \text{ impair : 1 racine : } \omega_0 = -V(-C, p)$$

3°) Relations de récurrence

$$\boxed{\text{R 1}} \quad T_p(\omega) T_q(\omega) = \frac{1}{2} [T_{p+q}(\omega) + T_{p-q}(\omega)]$$

$$\boxed{\text{R 2}} \quad T_p(T_q(\omega)) = T_q(T_p(\omega)) = T_{pq}(\omega)$$

$$\boxed{\text{R 3}} \quad T_{p+1}(\omega) = 2\omega T_p(\omega) - T_{p-1}(\omega)$$

$$\text{avec } T_0(\omega) = 1$$

$$T_1(\omega) = \omega$$

$$\boxed{\text{R 4}} \quad T_{2p}(\omega) = 2T_p^2(\omega) - 1.$$

4°) Propriété fondamentale

Si $\alpha > 1$, $\frac{T_p(\alpha x)}{T_p(\alpha)}$ est, parmi tous les polynômes de degré p et égaux à 1 pour $x=1$, celui qui est enserré sur $[-\alpha, +\alpha]$ entre les limites les plus étroites $\pm \frac{1}{T_p(\alpha)}$.

Remarque :

De nombreuses autres propriétés, relatives au polynôme de Tchebycheff (orthogonalité par exemple), ne sont pas citées ici, n'ayant aucun rapport avec le problème qui nous préoccupe.

(Se reporter par exemple à la référence [6]).

1. 1. 5. 4. Polynômes d'optimisation dans le cas général.

Le polynôme d'optimisation utilisant les propriétés fondamentales des polynômes de Tchebycheff, minimisant donc $\text{Max}_{x \in [\beta, \alpha]} |P_p(x)|$ a été donné par W. Markoff en 1892.

$$p_p(x) = \frac{T_p\left(\frac{-2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta}\right)}{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)} \quad (X - 1)$$

Ce polynôme vérifie bien les conditions exigées pour la convergence, car

$$p_p(0) = \frac{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)}{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)} = 1$$

$$\text{et } h = \text{Max}_{x \in [\beta, \alpha]} |p_p(x)| = \frac{1}{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)} < 1$$

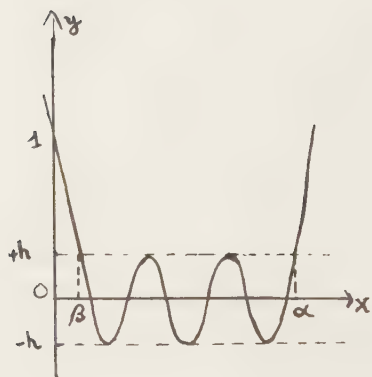
$$\text{car } \frac{\alpha + \beta}{\alpha - \beta} > 1 \implies T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right) = \text{chp}\left(\text{Argch}\left(\frac{\alpha + \beta}{\alpha - \beta}\right)\right) > 1.$$

1°) Variations

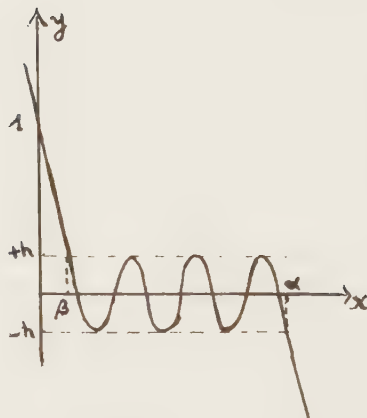
La courbe représentative de $p_p(x)$ prend l'un des deux aspects ci-dessous, suivant la parité de p . Les variations se déduisent de celles de $T_p(x)$. L'intervalle $[-1, +1]$ où $T_p(x)$ oscillait entre -1 et $+1$ est ramené à $[\beta, \alpha]$ par le changement

$$X = \frac{\alpha + \beta}{2} + \frac{\alpha - \beta}{2} x$$

$$\text{et dans cet intervalle, } p_p(x) \text{ oscille entre } +h \text{ et } -h \text{ avec } h = \frac{1}{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)}$$



$\frac{p \text{ pair}}{(p=6)}$



$\frac{p \text{ impair}}{(p=7)}$

Le polynôme $p_p(x)$ s'annule aux points d'abscisse

$$x_k = \frac{\alpha + \beta}{2} + \frac{\alpha - \beta}{2} \cos \frac{(2k-1)\pi}{2p} \quad k = 1, 2, \dots, p$$

et atteint h en valeur absolue aux points

$$x_i = \frac{\alpha + \beta}{2} + \frac{\alpha - \beta}{2} \cos \frac{i\pi}{p} \quad i = 0, 1, \dots, p$$

En particulier, ces valeurs sont atteintes aux points α et β

$$\left\{ \begin{array}{l} p_p(\alpha) = (-1)^p h \\ p_p(\beta) = h \end{array} \right. \quad \forall p$$

$$2^{\circ}) \text{ Etude de } h = \max_{x \in [\beta, \alpha]} |p_p(x)|$$

Nous savons que

$$h = \frac{1}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta} \right)} = \frac{1}{\operatorname{ch} p \operatorname{Argt} \operatorname{ch} z}$$

$$\text{avec } z = \frac{\alpha + \beta}{\alpha - \beta} = \frac{\alpha/\beta + 1}{\alpha/\beta - 1} = \frac{y + 1}{y - 1}$$

Mis sous cette forme, nous voyons que h est fonction de p et du rapport

$$\boxed{y = \alpha/\beta.}$$

Les bornes α et β n'interviennent donc, dans l'expression de h , que par leur rapport.

Etudions le comportement de h vis à vis de ces deux variables et les conséquences qui en découlent.

1. 1. 5. 5. Variations de $h = \max_{x \in [\beta, \alpha]} |p_p(x)|$ en fonction de $y = \alpha/\beta$

Supposons le degré fixé p , et étudions sommairement le comportement de h , lorsque y varie de 1 à l'infini, domaine qui nous intéresse.

$$h(y) = \frac{1}{T_p \left(\frac{y+1}{y-1} \right)} = \frac{1}{\operatorname{ch} p \left(\operatorname{Argt} \operatorname{ch} \left(\frac{y+1}{y-1} \right) \right)}$$

On vérifie immédiatement que $h(y)$ est fonction croissante, sa dérivée étant constamment positive pour $y \in]1, +\infty[$.

$$\frac{dh}{dy} = \frac{2p}{(y-1)^2} \cdot \frac{\operatorname{sh} p \left(\operatorname{Argt} \operatorname{ch} \left(\frac{y+1}{y-1} \right) \right)}{\operatorname{ch}^2 p \left(\operatorname{Argt} \operatorname{ch} \left(\frac{y+1}{y-1} \right) \right)} \cdot \frac{1}{\sqrt{\left(\frac{y+1}{y-1} \right)^2 - 1}}$$

Si $y = 1 + \varepsilon$ et $\varepsilon \rightarrow 0^+$

$$\left(\frac{dh}{dy} \right)_1 \simeq \lim_{\varepsilon \rightarrow 0^+} \left(\frac{2p}{\varepsilon \cdot e^{p \operatorname{Argt} \operatorname{ch}(2/\varepsilon)}} \right) = 0$$

Lorsque $y \rightarrow \infty$, $h(y)$ tend asymptotiquement vers 1 par valeurs inférieures, D'où la courbe grossière de variations :



Conséquences

Revenons au problème qui nous préoccupe : la résolution d'un système linéaire.

Les nombres α et β ont été définis respectivement comme majorante et minorante des valeurs propres λ_i .

$$\beta \leq \lambda_{\min} < \dots < \lambda_i$$

$$\lambda_i < \dots \leq \lambda_{\max} \leq \alpha$$

Donc :

$$\frac{\alpha}{\beta} \geq \frac{\lambda_M}{\lambda_m} \quad \text{si} \quad \left\{ \begin{array}{l} \lambda_M = \lambda_{\max} \\ \lambda_m = \lambda_{\min} \end{array} \right.$$

Il résulte de l'étude précédente que les nombres α et β devront être calculés de façon à rendre le rapport $y = \frac{\alpha}{\beta}$ le plus proche possible (par valeur supérieure) du rapport $y_0 = \frac{\lambda_M}{\lambda_m}$ pour lequel, p étant fixé, la convergence est la plus rapide, car $h(y_0) < h(y)$.

Les valeurs propres extrêmes devront donc être approchées, par procédé numérique puisqu'elles ne sont pas supposées strictement connues, avec le plus de précision possible. Un chapitre ultérieur sera consacré à cette détermination des deux nombres α et β .

Exemple numérique

$$\text{Supposons } \left| \begin{array}{l} \lambda_m = 1 \\ \lambda_M = 100 \end{array} \right. \implies y_0 = 100$$

et imposons un degré $p = 10$.

$$\begin{array}{ll} y_0 = 100 & h_0 = h(10, y_0) = 0,2641 \\ y_1 = 105 & h_1 = h(10, y_1) = 0,2767 \\ y_2 = 110 & h_2 = h(10, y_2) = 0,289 \\ y_3 = 120 & h_3 = h(10, y_3) = 0,3125 \\ & \left\{ \begin{array}{l} h_0^{14} = 0,802 \cdot 10^{-8} \\ h_2^{15} = 0,8214 \cdot 10^{-8} \\ h_3^{16} = 0,8284 \cdot 10^{-8} \end{array} \right. \end{array}$$

Ces trois dernières égalités montrent que 14 itérations correspondant à y_0 donnent une précision un peu supérieure à celle obtenue en 15 itérations ($y = y_2$), et 16 itérations ($y = y_3$).

1. 1. 5. 6. Variations de $h_p = \max_{x \in [\beta, \alpha]} |p_p(x)|$ en fonction du degré p

$$h(p) = \frac{1}{T_p\left(\frac{y+1}{y-1}\right)} = \frac{1}{\operatorname{ch} p \omega} \quad \text{avec } \omega = \operatorname{Arg}^t \operatorname{ch} \left(\frac{y+1}{y-1} \right)$$

Pour y fixé, donc ω , h décroît comme l'inverse de $\operatorname{ch} p \omega$. Sa décroissance est donc très rapide.

Elle est telle que l'on peut démontrer à l'aide de la relation

$$\boxed{h(q+p) < h(p) \cdot h(q)}$$

$$(\text{car } \frac{1}{\operatorname{ch}(p+q)(\omega)} = \frac{1}{\operatorname{ch} p \omega \cdot \operatorname{ch} q \omega + \operatorname{sh} p \omega \cdot \operatorname{sh} q \omega} < \frac{1}{\operatorname{ch} p \omega \cdot \operatorname{ch} q \omega})$$

que $h(q) < h(q_1) \cdot h(q_2) \cdot h(q_3) \cdot h(q_n)$ avec $\sum q_i = q$, ou que :

$$h(r * q) < h^q(r)$$

$$h(r * q) < h^r(q)$$

Une itération (III - 1) de degré $r * q$ optimise plus la convergence que q itérations (III - 1) de degré r , ou r itérations de degré q .

La relation R4 (chapitre I, 1.3.5) permet ainsi de comparer $h(2 * q)$ avec $h_{(q)}^2$

$$h_{2q} = \frac{h_q^2}{2 - h_q^2} \quad (XI - 1)$$

Exemple numérique

$$y = y_0 = 100$$

$p = 10$	$h(10) = 0,2641$	$h^2(10) = 0,0697$	$h^4(10) = 0,00486$	$h^8(10) = 0,0000236$
$p = 20$		$h(20) = 0,0361$	$h^2(20) = 0,0013$	$h^4(20) = 0,0000017$
$p = 40$			$h(40) = 0,000653$	$h^2(40) = 0,000000427$
$p = 80$				$h(80) = 0,000000213$

1. 1. 6. Résolution d'un système linéaire dans le cas général.

1. 1. 6. 1. La matrice $[A]$ est symétrique donc a toutes ses valeurs propres réelles.

Le théorème fondamental, tel qu'il a été démontré (1.1.3.4.) n'est valable que lorsque la matrice est diagonalisable et à valeurs propres toutes réelles.

$[A]$ peut donc être réelle symétrique ou symétrisable. Trois cas peuvent se rencontrer selon le signe des valeurs propres :

- les valeurs propres sont toutes positives
- les valeurs propres sont toutes négatives
- les valeurs propres ne sont pas toutes de même signe.

1°) *Valeurs propres toutes positives et localisées.*

C'est ce cas qui a été longuement traité dans ce chapitre. Rappelons que si $0 < \beta \leq \lambda_m \leq \dots \leq \lambda_i \leq \dots \leq \lambda_M \leq \alpha$

Le polynôme est :

$$p_p(x) = \frac{T_p \left(\frac{-2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right)}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta} \right)} \Rightarrow h_p = \frac{1}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta} \right)} (X - 1)$$

2°) Valeurs propres toutes négatives et localisées.

Supposons connus encore 2 nombres β et α , définis par

$$\alpha \leq \lambda_m < \dots < \lambda_i < \dots \lambda_M \leq \beta < 0.$$

$$\left| \begin{array}{l} |\beta| \leq \min_i |\lambda_i| \\ |\alpha| \geq \max_i |\lambda_i| \end{array} \right|$$

On démontre facilement que le polynôme à choisir est encore

$$p_p(x) = \frac{T_p \left(\frac{-2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right)}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta} \right)}$$

avec

$$h_p = \frac{1}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta} \right)}$$

3°) Les valeurs propres de $[A]$ sont réelles mais ne sont pas toutes de même signe.

Nous pouvons supposer ces valeurs propres localisées dans un intervalle $[b, a]$ avec

$$\left\{ \begin{array}{l} b \text{ négatif} \leq \lambda_i \quad (i=1, \dots, n) \\ \alpha \text{ positif} \geq \lambda_i \quad (i=1, \dots, n) \end{array} \right.$$

$$\Rightarrow b \leq \lambda_m < \lambda_2 < \dots \leq \lambda_k < 0 < \lambda_{k+1} \dots \leq \lambda_j \dots \leq \lambda_{n-1} \leq \lambda_M \leq \alpha$$

La matrice $[A]$ étant inversible n'a pas de valeurs propres nulles, il existe donc un nombre c défini par $c \leq \min_i |\lambda_i|$.

A ce niveau, deux possibilités s'offrent :

a) remplacer le système par le système équivalent

$$[A^t][A][X] = [A^t][B]$$

système défini positif, et à valeurs propres réelles toutes positives.

b) rechercher un polynôme $p(x)$ tel que $\text{Max } |p(x)|$ soit minimum sur

$$[b, -|c|] \cup [|c|, a] \text{ et } p(0) = 1.$$

Ⓐ $[A][X] = [B] \iff [A^t][A][X] = [A^t][B] = [C]$

La matrice $[A] = [A^t][A]$ du nouveau système est définie positive, et a ses valeurs propres, toutes positives $\forall i \lambda_i^* \in [\beta, \alpha]$

avec :

$$\alpha = \begin{cases} a^2 & \text{si } |a| > |b| \\ b^2 & \text{si } |a| < |b| \end{cases}$$

et $\beta = c^2$

$$\implies p_p(x) = \frac{T_p\left(\frac{-2x}{a^2 - c^2} + \frac{a^2 + c^2}{a^2 - c^2}\right)}{T_p\left(\frac{a^2 + c^2}{a^2 - c^2}\right)} \text{ ou } p_p(x) = \frac{T_p\left(\frac{-2x}{b^2 - c^2} + \frac{b^2 + c^2}{b^2 - c^2}\right)}{T_p\left(\frac{b^2 + c^2}{b^2 - c^2}\right)}$$

si $|a| > |b|$ $|a| < |b|$

(XI - 1)

alors $h_p = \frac{1}{T_p\left(\frac{a^2 + c^2}{a^2 - c^2}\right)}$ ou $h_p = \frac{1}{T_p\left(\frac{b^2 + c^2}{b^2 - c^2}\right)}$

selon que $|a| > |b|$ ou $|b| > |a|$.

Ⓑ Si, pour une raison quelconque, nous ne pouvons pas transformer le système $[A][X] = [B]$ en $[A^t][A][X] = [A^t][B]$, il est possible de trouver un polynôme vérifiant les conditions du théorème (1.1.3.4.), tel que $p_p(0) = 1$, et tel que $\text{Max}_x |p_p(x)|$ soit minimum sur $[b, -c] \cup [c, a]$. Nous modifions légèrement cet intervalle et recherchons $p(x)$ tel que $\text{Max}_x |p_p(x)|$ soit minimum sur $[b, -c] \cup [c, |b|]$ si $|b| > a$ ou sur $[-a - c] \cup [c, a]$ si $|b| < a$

Ce polynôme, symétrique par rapport à Oy, donc pair, se détermine aisément à partir de l'expression générale de $P_p(x)$. Il vaut ici :

$$p_{2p}^*(x) = T_p\left(\frac{-2x^2}{b^2 - c^2} + \frac{b^2 + c^2}{b^2 - c^2}\right) \bigg/ T_p\left(\frac{b^2 + c^2}{b^2 - c^2}\right) \quad \text{si } |b| > a$$

$$p_{2p}^*(x) = T_p\left(\frac{-2x^2}{a^2 - c^2} + \frac{a^2 + c^2}{a^2 - c^2}\right) \bigg/ T_p\left(\frac{a^2 + c^2}{a^2 - c^2}\right) \quad \text{si } |b| < a$$

$$h_{2p}^* = \frac{1}{T_p\left(\frac{a^2 + c^2}{a^2 - c^2}\right)} \quad \text{ou} \quad \frac{1}{T_p\left(\frac{b^2 + c^2}{b^2 - c^2}\right)}$$

selon que $|a| > |b|$ ou que $|b| > |a|$.

Remarque :

Ce polynôme de degré $2p$ assure une convergence aussi rapide (car $h_p = h_{2p}^*$) que le polynôme de degré p , correspondant à la résolution du système $[A^t][A][X] = [A^t][B]$. Il y a donc intérêt à transformer le système en se ramenant au système défini positif, la résolution étant deux fois plus rapide.

1. 1. 6. 2. Cas général

La matrice et le second membre peuvent être soit réels, soit à éléments complexes. Dans les deux cas, on se ramènera à un système équivalent défini positif à valeurs propres réelles et l'on adoptera le polynôme correspondant, déjà défini ($X - 1$).

1°) *La matrice et le second membre sont à éléments réels*

On transforme le système $[A][X] = [B]$, de N équations à N inconnues, pour se ramener au système de dimension équivalente $[\mathcal{A}][X] = [\mathcal{B}]$, système défini positif et à valeurs propres positives, avec $[\mathcal{A}] = [A^t][A]$ et $[\mathcal{B}] = [A^t][B]$.

2°) *La matrice et le second membre sont complexes*

Le système $[A][Z] = [B]$ peut s'écrire :

$$[[A_1] + i[A_2]] [[X_1] + i[X_2]] = [[B_1] + i[B_2]]$$

En isolant les parties réelles et les parties imaginaires, on peut se ramener à un système à éléments tous réels, de dimension $2n$, qui permet d'obtenir la solution : c'est le système :

$$\begin{bmatrix} [A_1] & -[A_2] \\ [A_2] & [A_1] \end{bmatrix} \begin{bmatrix} [X_1] \\ [X_2] \end{bmatrix} = \begin{bmatrix} [B_1] \\ [B_2] \end{bmatrix}$$

1. 2. TECHNIQUES

1. 2. 1. Introduction

Nous nous proposons de décrire dans ce chapitre différents algorithmes permettant le passage du vecteur $[X_k]$ au vecteur itéré $[X_{k+1}] = [P_p(A)] [X_k] - [Q_{p-1}(A)] [B]$, lorsque le polynôme $[P_p(A)]$ est défini par $(X - 1)$.

Il est possible dans le cas de petites matrices, d'explicitier directement les matrices $[P_p(A)]$ et $[Q_{p-1}(A)]$ et d'itérer ensuite la relation ci-dessus.

Nous nous limitons dans cette étude à une autre technique, où les matrices $[P_p(A)]$ et $[Q_{p-1}(A)]$ ne sont jamais explicitées mais où le vecteur $[P_p(A)] [X_k] - [Q_{p-1}(A)] [B]$ est calculé à chaque itération de (III - 1), que nous appellerons dorénavant "*itération principale*", par un procédé itératif nécessitant p produits du type $[A] [X]$.

Les itérations correspondantes, au nombre de p , seront, elles, appelées indifféremment "*itérations secondaires*" ou "*sous itérations*", pour les différencier des itérations principales.

La technique adoptée se prête particulièrement bien au traitement des grands systèmes linéaires "pleins", sur petits ordinateurs, équipés d'une mémoire extérieure (bande magnétique, disque magnétique ou tambour), où est stockée la matrice $[A]$.

De plus, lorsque la méthode étudiée est destinée à résoudre de très grands systèmes - aux différences finies par exemple -, la matrice $[A]$ est très creuse et n'est jamais explicitée, et il est facile d'écrire un algorithme effectuant le produit $[A] [X]$ en ne tenant compte que des éléments non nuls de $[A]$.

Plusieurs possibilités sont envisageables pour construire les vecteurs itérés principaux $[X_k]$.

Une première consiste à utiliser les coefficients du polynôme $[P_p(A)]$, (et ceux de $[Q_{p-1}(A)]$, qui s'en déduisent).

Elle a été rapidement abandonnée, car elle se révélait très instable : Elle ne permettait pas d'itérer des polynômes de degré supérieur à 10, ce qui supprime l'intérêt de la méthode.

Deux autres possibilités ont retenu notre attention et vont être l'objet de ce travail.

Elles reflètent deux façons de calculer un polynôme de Tchebycheff.

1°) à l'aide de ses racines, toutes connues :

$$T_p(\omega) = K \prod_{i=1}^p (\omega - \omega_i)$$

K est une constante qui disparaîtra lors du calcul de $[P_p(A)]$

2°) par utilisation de la relation de récurrence liant 3 polynômes de Tchebycheff de degrés consécutifs.

$$T_p(\omega) = 2\omega T_{p-1}(\omega) - T_{p-2}(\omega)$$

avec $T_0(\omega) = 1$

$$T_1(\omega) = \omega$$

Les deux techniques correspondantes, que nous appellerons technique RAC (racines) pour la première, et technique REC (recurrence) pour la seconde seront décrites, puis comparées à des méthodes connues : La méthode de Richardson, la méthode de Stiefel-Frank et la méthode semi-itérative de Tchebycheff.

1. 2. 2. Technique R.A.C. : Utilisation des racines du polynôme $p_p(x)$.

Comparaison avec la méthode de Richardson

1. 2. 2. 1. Description des sous itérations

Nous cherchons à calculer le vecteur itéré principal, $[X_{k+1}]$, fonction de $[X_k]$, en faisant intervenir les p racines du polynôme $p_p(x)$, associée à $[P_p(A)]$

$$p_p(x) = (1-r_1^{(p)} x) (1-r_2^{(p)} x) \dots (1-r_p^{(p)} x) = \prod_{i=1}^p (1-r_i^{(p)} x)$$

avec $r_i^{(p)} = \frac{1}{x_i^{(p)}}$

mis sous cette forme il apparait bien que $p_n(0) = 1$.

$$[P_p(A)] = [[I] - r_1^{(p)}[A]] [[I] - r_2^{(p)}[A]] \dots [I] - r_p^{(p)}[A]]$$

définissons p sous itérations $[Y_i^k] = f(Y_i^k) = f(Y_{i-1}^k, r_i^{(p)})$

$$\begin{aligned} [Y_1^k] &= [I] - r_1^{(p)}[A] \quad [Y_0^k] + r_1^{(p)}[B] \\ [Y_2^k] &= [I] - r_2^{(p)}[A] \quad [Y_1^k] + r_2^{(p)}[B] \\ &\dots \dots \dots (X-2) \\ [Y_i^k] &= [I] - r_i^{(p)}[A] \quad [Y_{i-1}^k] + r_i^{(p)}[B] \\ &\dots \dots \dots \\ [Y_p^k] &= [I] - r_p^{(p)}[A] \quad [Y_{p-1}^k] + r_p^{(p)}[B] \end{aligned}$$

on vérifie simplement que si $[X]$ est solution de $[A][X] = [B]$

$$[[Y_p^k] - [X]] = [P_p(A)] [[Y_0^k] - [X]]$$

donc puisque la polynôme $p_p(x)$ vérifie les conditions du théorème fondamental (1.1.3.4),

$$[Y_p^k] = [P_p(A)] [Y_0^k] - [Q_{p-1}(A)] [B]$$

Si $[Y_0^k] = [X_k]$, kième itéré de la suite (III - 1), les p " sous itérations" ($X - 2$), permettent d'obtenir $[X_{k+1}] = [Y_p^k]$

1. 2. 2. 2. Technique de Balayage systématique

Nous avons défini les $r_1^{(p)}$, paramètres de relaxation des sous itérations ($X - 2$), comme étant les inverses des racines du polynôme

$$p_p(x) = \frac{T_p\left(\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta}\right)}{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)}$$

ces racines sont définies par

$$\boxed{x_k^{(p)} = \frac{\alpha + \beta}{2} - \frac{\alpha - \beta}{2} \cos\left(\frac{(2k-1)\pi}{2p}\right)} \quad (XI - 2)$$

$$k = 1, 2, \dots, p$$

Lorsque k varie de 1 à p , les $x_k^{(p)}$ croissent de

$$x_1^{(p)} = \frac{\alpha + \beta}{2} - \frac{\alpha - \beta}{2} \cos\left(\frac{\pi}{2p}\right)$$

à

$$x_p^{(p)} = \frac{\alpha + \beta}{2} - \frac{\alpha - \beta}{2} \cos\left(\frac{(2p-1)\pi}{2p}\right)$$

Dans les premiers essais d'utilisation de la technique R.A.C. on employait les $r_i^{(p)}$ égaux à $\frac{1}{x_1^{(p)}}$, c'est-à-dire décroissants ($i = 1, \dots, p$).

Le processus se révélait alors très instable, et les degrés maximums, utilisables étaient faibles (≤ 15).

En associant à $r_i^{(p)}$ les valeurs $r_i^{(p)} = \frac{1}{x^{(p)}}$, c'est-à-dire en itérant (X - 2) avec des $r_i^{(p)}$ croissants, l'instabilité s'est révélée être encore plus intense.

Le choix de l'ordre des racines $x_k^{(p)}$ à associer aux paramètres $r_i^{(p)}$ était donc essentiel.

David Young (réf. n° 4) en utilisant la méthode de Richardson (voir ci-dessous 1.2.2.3.), s'était heurté au même phénomène. Une étude expérimentale a été faite à ce sujet, et l'ordre d'appel des racines, qui s'en déduit, a permis l'utilisation de polynômes de degrés élevés (240).

Balayage systématique - Justification et description

Essayons tout d'abord d'expliquer la cause de l'instabilité qui se manifestait lorsqu'on appelait les racines $x_i^{(p)}$ par ordre croissant. A cet effet introduisons la représentation $[A] = [\Omega] \begin{bmatrix} \dots & \lambda_i & \dots \end{bmatrix} [\Omega]^1$

La première sous itération de (X-2) peut alors être écrite :

$$[Y_1^{(k)}] = [\Omega] \begin{bmatrix} (1-r_1^{(p)}\lambda_1) & \text{O} & \text{O} \\ \text{O} & (1-r_1^{(p)}\lambda_2) & \\ \vdots & \vdots & \vdots \\ \text{O} & \text{O} & (1-r_1^{(p)}\lambda_n) \end{bmatrix} [\Omega]^{-1} [Y_0^{(k)}] + r_1^{(p)} [B]$$

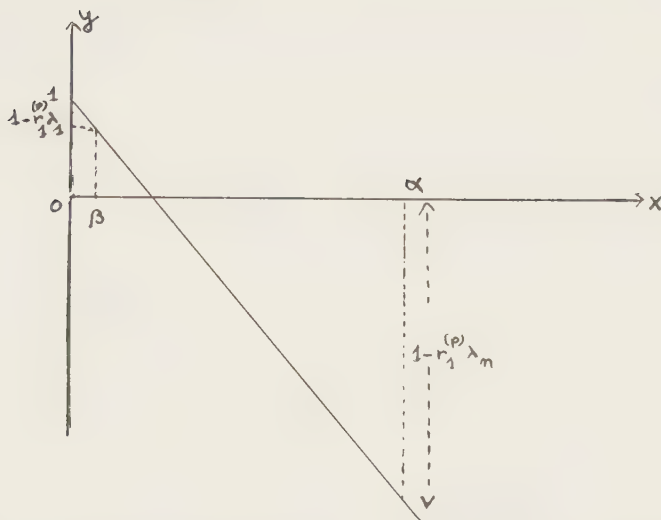
Si le degré p est élevé, la racine $x_1^{(p)}$ est très proche par valeur supérieure de β

$$x_1^{(p)} = \frac{\alpha + \beta}{2} - \frac{\alpha - \beta}{2} \left(1 - \frac{\pi}{2p}\right)$$

$z_1^{(p)} = 1 - r_1^{(p)}\lambda_1$ est très petit alors que $z_n^{(p)} = 1 - r_1^{(p)}\lambda_n$ est très grand.

Les autres $z_i^{(p)} = 1 - r_1^{(p)}\lambda_i$ vérifiant tous $0 < |z_1^{(p)}| < |z_n^{(p)}|$.

Il y a une disproportion importante entre les éléments les plus grands et les éléments les plus petits de la matrice diagonale.



associons alors à la deuxième sous-itération de $(X - 2)$, la valeur $r_2^{(p)} = \frac{1}{x_2^{(p)}}$

$$[Y_2^k] = [[I] - r_2^{(p)}[A]] [Y_1^k] + r_2^{(p)}[B]$$

$$[Y_2^k] = [[I] - r_2^{(p)}[A]] [[I] - r_1^{(p)}[A]] [Y_0^k] + [[I] - r_2^{(p)}[A]] r_1^{(p)}[B] + r_2^{(p)}[B]$$

$$[Y_2^k] = [\Omega] \left[\dots (1 - r_1^{(p)} \lambda_1) (1 - r_2^{(p)} \lambda_1) \dots \right] [\Omega^{-1}] [Y_0^k] + f(A, r_1^{(p)}, r_2^{(p)}, B)$$

dans la matrice diagonale.

$$\left[\dots (1 - r_1^{(p)} \lambda_1) (1 - r_2^{(p)} \lambda_1) \dots \right] = \begin{bmatrix} (1 - r_1^{(p)} \lambda_1) & (1 - r_2^{(p)} \lambda_1) & \dots & 0 \\ 0 & & & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & & 0 \\ 0 & \dots & 0 & (1 - r_1^{(p)} \lambda_n) (1 - r_2^{(p)} \lambda_n) \end{bmatrix}$$

, la disproportion entre $(1 - r_1^{(p)} \lambda_1) (1 - r_2^{(p)} \lambda_1)$ et $(1 - r_1^{(p)} \lambda_n) (1 - r_2^{(p)} \lambda_n)$ est à peu près le carré de celle qui intervenait entre $z_1^{(p)}$ et $z_n^{(p)}$ lors de la première sous-itération. Cette disproportion augmentera ainsi jusqu'à la valeur $r_i^{(p)} \approx \frac{2}{\alpha + \beta}$, pour théoriquement décroître, puis disparaître avec $r_i^{(p)} = \frac{1}{x_p^{(p)}}$.

Or les calculateurs effectuent les opérations arithmétiques avec un nombre de chiffres significatifs fixe, et les troncatures qui se produisent à chaque opération, peuvent par cumul, conduire à des résultats aberrants. Cet effet est plus important lorsque les opérations portent sur des nombres de grandeur non comparable, ce qui se produit lors des premières sous-itérations.

Ces quelques considérations très hypothétiques, car nous sommes très mal armés pour étudier l'influence des erreurs de troncature, ont guidé notre choix dans l'ordre d'utilisation des racines. Le choix que nous avons adopté repose sur le principe suivant :

Nous cherchons à chaque "sous-itération" de $(X - 2)$, à réduire au mieux la disproportion qui a pu se produire, lors de l'étape précédente.

Ainsi, à l'itération secondaire n° 1 est associé $r_1^{(p)} = \frac{1}{x_1^{(p)}}$

, à l'itération secondaire n° 2 $r_2^{(p)} = \frac{1}{x_p^{(p)}}$

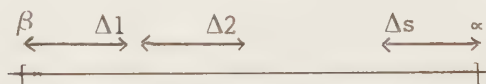
à ce niveau, la plus grande valeur est $(1 - r_1^{(p)} \lambda_1) (1 - r_2^{(p)} \lambda_1)$, où λ_1 est la valeur propre la plus proche de $\frac{\alpha + \beta}{2}$

d'où $r_3^{(p)} = \frac{1}{x_{p/2}^{(p)}}$ (si p est pair)

, et ainsi de suite ...

A cet ordonnancement des racines, nous avons associé un balayage, proprement dit.

Partitionnons $[\beta, \alpha]$ en s intervalles groupant chacun q racines consécutives de $p_p(x)$ avec $p = q * s$



avec

$$\begin{array}{l} (x_1^{(p)}, x_2^{(p)}, \dots, x_q^{(p)}) \in \Delta 1 \\ (x_{q+1}^{(p)}, x_{q+2}^{(p)}, \dots, x_{2q}^{(p)}) \in \Delta 2 \\ \dots \dots \dots \\ (x_{(s-1)q+1}^{(p)}, \dots, x_{qs}^{(p)}) \in \Delta s \end{array}$$

définissons q polynômes $r_i^s(x)$, de degré s, s'annulant chacun une fois dans tous les Δ_i , pour une des racines de $p_p(x)$ appartenant à cette intervalle.

$$p_p(x) = r_1^s(x) \dots r_2^s(x) \dots r_q^s(x)$$

Les "sous-itérations" ($X - 2$), ont été décomposées en q groupes de s "sous-itérations", chacun de ces groupes utilisant les s racines du polynôme $i_1^s(x)$ correspondant, l'appel de ses racines s'effectuant suivant la règle édictée ci-dessus.

Ce balayage a été utilisé dans le seul souci de retrouver périodiquement une racine de chaque Δ_i .

Le balayage systématique adopté, synthétisant les idées décrites ci-dessus, et qui a donné les résultats les plus satisfaisants peut être explicité par l'instruction ALGOL ci-dessous.

```
POUR I := 1 PAS 1 JUSQUA Q FAIRE
POUR J := 1 PAS 1 JUSQUA S/4 FAIRE
POUR K := I + (J - 1) * Q, P - I + 1 - (J - 1) * Q,
    P/2 - I - (J - 1) * Q + 1, P/2 + I + (J - 1) * Q FAIRE

    DEBUT    U := (BETA + ALPHA)/2 - (ALPHA - BETA)/2
              * COS ((2 * K - 1) * 3.1415926/2/P) ;
    MULMAVEC (X, Y, N) ;
    POUR L := 1 PAS 1 JUSQUA N FAIRE
        Y [L] := X [L] - Y [L] / U + B [L] / U

FIN ;
```

NB. : Ce balayage adopté n'exclut pas d'autres tout aussi valables, et il est possible qu'il en existe de meilleurs.

Remarques : Q et S ont les significations décrites par ailleurs

$$P = Q * S \quad \text{avec } S \text{ divisible par } 4$$

$$ALPHA = \max_i (\lambda_i)$$

$$BETA = \min_i (\lambda_i)$$

MULMAVEC (X, Y, N) est la procédure qui exécute le produit $[A] [X]$ et place le produit dans le tableau $[Y]$, N étant la dimension de $[A]$

à chaque boucle. U représente la racine $x_i^{(p)}$ appelée.

Exemple $P = 24$ avec $Q = 3$ et $S = 8$

Si $x_1, x_2, \dots, x_{24}$ représentent les racines classées par ordre croissant, l'ordre d'appel au cours des itérations (X - 2), sera le suivant.

$$\begin{array}{|c|c|} \hline x_1, & x_{24}, & x_{12}, & x_{13}, \\ \hline x_2, & x_{23}, & x_{11}, & x_{14}, \\ \hline x_3, & x_{22}, & x_{10}, & x_{15}, \\ \hline \end{array} \quad \begin{array}{|c|c|} \hline x_4, & x_{21}, & x_9, & x_{16}, \\ \hline x_5, & x_{20}, & x_8, & x_{17}, \\ \hline x_6, & x_{19}, & x_7, & x_{18}, \\ \hline \end{array}$$

1. 2. 2. 3. Comparaison avec la Méthode itérative de Richardson.

1°) Méthode de Richardson - Historique et description

La méthode, qui porte son nom a été présentée en 1910 par L. F. Richardson. Son but était alors de résoudre des systèmes d'équations linéaires intervenant dans des problèmes aux différences finies associées à des équations aux dérivées partielles elliptiques.

La formule itérative qu'il proposait, est linéaire et du 1er degré, c'est-à-dire fonction linéaire de l'itérè précédent.

$$[X_i] = [X_{i-1}] - r_i[A] [X_{i-1}] + r_i[B] \quad (\text{XII} - 2)$$

avec $[X_0]$ donné

au bout de N itérations, le vecteur $[X_N]$ est tel que :

$$[[X_N] - [X]] = [[I] - r_1[A]] [[I] - r_2[A]] \dots [[I] - r_N[A]] [[X_0] - [X]]$$

si $[X]$ est la solution de $[A] [X] = [B]$

Richardson donnait quelques indications suggestives, pour rendre $[X_N]$ le plus proche possible de la solution et en fait posait le problème du choix des r_i .

Ce choix ($r_i = \frac{1}{x_i}$, avec x_i racine de $T_n(\frac{-2x + \alpha + \beta}{\alpha + \beta})$) a été imposé par les travaux de Markoff.

Il a fallu attendre l'avènement des calculatrices électroniques, pour permettre, vers les années 1950, l'expérimentation de cette méthode à des systèmes définis positifs, entre autres, par David Young (réf. n° 4), R.S. Varga (réf. n° 3), et W.L. Frank (réf. n° 7).

2°) Comparaison technique RAC - Méthode de Richardson

Au vu des itérations (XII - 2) proposées par Richardson, et des "sous-itérations" (X - 2), introduites dans la technique R.A.C., le lien qui unit ces deux méthodes semble évident.

Mais alors que les *itérations de Richardson* devaient dans son esprit, permettre d'atteindre directement la solution approchée voulue, les "*sous-itérations*" (X - 2) ne jouent elles qu'un rôle d'intermédiaire permettant le passage de l'itéré $[X_k]$ à l'itéré "principal" suivant $[X_{k+1}]$ de la suite principale (III - 1).

En d'autres termes, Richardson veut atteindre la solution désirée $[X^*]$ en une itération principale de degré N,

$$\implies [X^*] = [P_N(A)] [X_0] - [Q_{N-1}(A)] [B]$$

alors que nous nous proposons d'atteindre une solution équivalente $[X^{**}]$ en K itérations principales de degré P.

$$[X^{**}] = [X_K] \text{ avec}$$

$$[X_i] = [P_p(A)] [X_{i-1}] - [Q_{p-1}(A)] [B]$$

La méthode de Richardson, nécessite, avant de débiter les itérations, la connaissance du nombre N, car chaque r_i en dépend.

Ce nombre, permettant l'obtention de la précision voulue peut se déterminer de façon théorique, en fonction de $\epsilon = \frac{\| [X] - [X^*] \|}{\| [X] - [X_0] \|}$ et du rapport $\frac{\alpha}{\beta}$

Le cumul des erreurs de troncature, qui se produisent durant les calculs, peut rendre le résultat final, très éloigné de la solution voulue. *La méthode de Richardson est alors à rejeter.*

Un autre des inconvénients majeurs qu'elle présente, est qu'aucune possibilité n'est offerte, pour tester la bonne marche du processus itératif, durant son déroulement.

En introduisant deux niveaux d'itérations, comme nous le proposons, les itérations principales, et les sous-itérations, nous pouvons périodiquement contrôler la bonne marche de la convergence en faisant appel aux formules d'erreur explicitées en 1.1.4, la formule la plus intéressante étant

$$\frac{\| [X_i] - [X_{i-1}] \|}{\| [X_{i-1}] - [X_{i-2}] \|} \leq h < 1 \quad (\text{si } [A] \text{ est définie positive}).$$

Nous reviendrons, plus en détail, par la suite, sur ces notions de contrôle d'erreur, et sur la politique de convergence que nous avons adoptée.

1. 2. 3. Technique R. E. C. : Utilisation de la relation de récurrence liant trois polynômes consécutifs de Tchebycheff. Comparaison avec la méthode de Stiefel-Frank.

1. 2. 3. 1. Description de la Technique R. E. C.

Pour passer du vecteur $[X_k]$ au vecteur itéré suivant de la suite principale $[X_{k+1}] = [P_p(A)] [X_k] - [Q_{p-1}(A)] [B]$, nous introduisons p vecteurs intermédiaires définis par

$$\begin{array}{l} [Y_1^k] = [P_1(A)] [Y_0^k] - [Q_0(A)] [B] \\ [Y_2^k] = [P_2(A)] [Y_0^k] - [Q_1(A)] [B] \\ \text{-----} \\ [Y_i^k] = [P_i(A)] [Y_0^k] - [Q_{i-1}(A)] [B] \\ \text{-----} \\ [Y_p^k] = [P_p(A)] [Y_0^k] - [Q_{p-1}(A)] [B] \end{array} \quad (\text{XIII} - 2)$$

Si $[Y_0^k] = [X_k]$, nous voyons que $[Y_p^k]$ s'identifie à $[X_{k+1}]$.

La relation de récurrence (R3) liant trois polynômes de Tchebycheff de degrés successifs, va nous permettre de relier trois vecteurs $[Y_i^k]$, $[Y_{i-1}^k]$ et $[Y_{i-2}^k]$. Le calcul de $[X_{k+1}] = [Y_p^k]$ pourra ainsi se faire de proche en proche, à partir de la connaissance de $[Y_0^k] = [X_k]$ et $[Y_1^k]$.

1°) Relation de récurrence liant trois polynômes $P_i(A)$ de degrés successifs

Cette relation s'établit à partir de la formule

$$\text{- R3 -} \quad T_i(\omega) = 2 \omega T_{i-1}(\omega) - T_{i-2}(\omega)$$

$$\text{avec} \quad \begin{cases} T_0(\omega) = 1 \\ T_1(\omega) = (\omega) . \end{cases}$$

$$p_i(x) = \frac{T_i\left(\frac{-2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta}\right)}{T_i\left(\frac{\alpha + \beta}{\alpha - \beta}\right)}$$

$$C_i P_i(x) = T_i \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right) \quad (*)$$

$$C_i = T_i \left(\frac{\alpha + \beta}{\alpha - \beta} \right)$$

$$R3 \Rightarrow \boxed{C_i = 2 \left(\frac{\alpha + \beta}{\alpha - \beta} \right) C_{i-1} - C_{i-2}} \quad (XIV - 2)$$

avec $\left| \begin{array}{l} C_0 = 1 \\ C_1 = \frac{\alpha + \beta}{\alpha - \beta} \end{array} \right.$

$$(*) \quad T_i \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right) = 2 \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right) T_{i-1} \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right) - T_{i-2} \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right)$$

Il s'en suit

$$P_i(x) = 2 \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right) \frac{C_{i-1}}{C_i} P_{i-1}(x) - \frac{C_{i-2}}{C_i} P_{i-2}(x)$$

$$\text{avec } \begin{array}{l} P_0(x) = 1 \\ P_1(x) = T_1 \left(-\frac{2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} \right) / C_1 = -\frac{2x}{\alpha + \beta} + 1 \end{array}$$

D'où la relation liant trois polynômes matriciels $[P_i(A)]$, $[P_{i-1}(A)]$, et $[P_{i-2}(A)]$

$$\boxed{[P_i(A)] = 2 \frac{C_{i-1}}{C_i} \left[-\frac{2[A]}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} [I] \right] [P_{i-1}(A)] - \frac{C_{i-2}}{C_i} [P_{i-2}(A)]}$$

(XV - 2) et $[P_0(A)] = [I]$

$$[P_1(A)] = [I] - \frac{2}{\alpha + \beta} [A]$$

2°) *Itérations secondaires de la technique R. E. C.*

Reportons la valeur (XV - 2) de $[P_i(A)]$ dans la sous-itération correspondante de (XIII - 2).

$$[Y_i^k] = [P_i(A)] [Y_0^k] - [Q_{i-1}(A)] [B]$$

$$[Y_i^k] = 2 \frac{C_{i-1}}{C_i} \left[-\frac{2[A]}{\alpha-\beta} + \frac{\alpha+\beta}{\alpha-\beta} [I] \right] [P_{i-1}(A)] [Y_0^k] - \frac{C_{i-2}}{C_i} [P_{i-2}(A)] [Y_0^k] - [Q_{i-1}(A)] [B]$$

$$(XIII - 2) \Rightarrow \begin{cases} [P_{i-1}(A)] [Y_0^k] = [Y_{i-1}^k] + [Q_{i-2}(A)] [B] \\ [P_{i-2}(A)] [Y_0^k] = [Y_{i-2}^k] + [Q_{i-3}(A)] [B] \end{cases}$$

$$[Y_i^k] = 2 \frac{C_{i-1}}{C_i} \left[-\frac{2[A]}{\alpha-\beta} + \frac{\alpha+\beta}{\alpha-\beta} [I] \right] [Y_{i-1}^k] - \frac{C_{i-2}}{C_i} [Y_{i-2}^k] + [R(A)] [B]$$

$$\text{avec } [R(A)] = 2 \frac{C_{i-1}}{C_i} \left[\left(-\frac{2[A]}{\alpha-\beta} + \frac{\alpha+\beta}{\alpha-\beta} [I] \right) [Q_{i-2}(A)] - \frac{C_{i-2}}{C_i} [Q_{i-3}(A)] - [Q_{i-1}(A)] \right]$$

$$[Q_{j-1}(A)] = [[P_j(A)] - [I]] [A^{-1}]$$

$$[A] [R(A)] = [S(A)] + [T(A)]$$

$$\begin{cases} [S(A)] = 2 \frac{C_{i-1}}{C_i} \left[\left(-\frac{2[A]}{\alpha-\beta} + \frac{\alpha+\beta}{\alpha-\beta} [I] \right) [P_{i-1}(A)] - \frac{C_{i-2}}{C_i} [P_{i-2}(A)] - [P_i(A)] \right] \\ [T(A)] = 4 \frac{C_{i-1}[A]}{C_i \alpha - \beta} - \left(2 \frac{C_{i-1}}{C_i} \frac{\alpha + \beta}{\alpha - \beta} - \frac{C_{i-2}}{C_i} - 1 \right) [I] \end{cases}$$

$$(XV - 2) \text{ et } (XIV - 2) \Rightarrow \begin{cases} [S(A)] = [0] \\ [T(A)] = 4 \frac{C_{i-1}}{C_i} \frac{[A]}{\alpha - \beta} \end{cases}$$

$$\Rightarrow [R(A)] = 4 \frac{C_{i-1}}{C_i} \frac{[I]}{\alpha - \beta}$$

Les p itérations secondaires, du second degré, permettant le passage de $[X_k]$ à $[X_{k+1}] = [Y_p^k]$ sont donc définies par

$$[Y_i^k] = 2 \frac{C_{i-1}}{C_i} \left[-\frac{2[A]}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta} [I] \right] [Y_{i-1}^k] - \frac{C_{i-2}}{C_i} [Y_{i-2}^k] + \frac{4}{\alpha - \beta} \frac{C_{i-1}}{C_i} [B]$$

$$(XVI - 2) \quad \text{avec} \quad \begin{cases} [Y_0^k] = [X_k] \\ [Y_1^k] = \left[[I] - \frac{2[A]}{\alpha + \beta} \right] [X_k] + \frac{2}{\alpha + \beta} [B] \end{cases}$$

les C_i étant définis par (XIV - 2)

Remarque : (XIV - 2)
$$2 \frac{C_{i-1}}{C_i} \frac{\alpha + \beta}{\alpha - \beta} = 1 + \frac{C_{i-2}}{C_i}$$

Les sous-itérations (XVI - 2) peuvent prendre la forme équivalente

$$[Y_i^k] = [Y_{i-1}^k] + \frac{C_{i-2}}{C_i} [[Y_{i-1}^k] - [Y_{i-2}^k]] + \frac{4}{\alpha - \beta} \frac{C_{i-1}}{C_i} [[B] - [A] [Y_{i-1}^k]]$$

$$(XVII - 2) \quad \text{avec} \quad \begin{cases} [Y_0^k] = [X_k] \\ [Y_1^k] = \left[[I] - \frac{2[A]}{\alpha + \beta} \right] [X_k] + \frac{2}{\alpha + \beta} [B] \end{cases}$$

1. 2. 3. 2. Méthode de Stiefel-Frank - Comparaison avec la Technique R.E.C.

1°) Description

L'importance des erreurs de troncature rencontrées lors de l'expérimentation de la Méthode de Richardson, et l'instabilité qui en découlait parfois, ont souvent rendu cette méthode sans intérêt pratique.

Certains auteurs, et en particulier Stiefel (réf. 5) et Werner L. Frank (réf. 7), ont imaginé d'utiliser la récurrence (R 3) liant trois polynômes de Tchebycheff de degrés consécutifs.

L'itération proposée de la méthode, que nous pouvons appeler *Méthode de Stiefel Frank* est une itération liant trois itères successifs

$$[X_k] = \frac{4}{\alpha - \beta} \frac{C_{k-1}}{C_k} [[B] - [A] [X_{k-1}]] + \frac{C_{k-2}}{C_k} [[X_{k-1}] - [X_{k-2}]] + [X_{k-1}]$$

pour $k \geq 1$

$$(XVIII - 2) \quad [X_1] = [X_0] + \frac{2}{\alpha + \beta} [[B] - [A] [X_0]]$$

avec $C_0 = 1$

$$C_1 = \frac{\alpha + \beta}{\alpha - \beta}$$

et $C_k = 2 \frac{\alpha + \beta}{\alpha - \beta} C_{k-1} - C_{k-2}$

$k \geq 1$

Le vecteur $[X_N]$ ainsi obtenu à la $N^{\text{ième}}$ itération de (XVIII - 2) est théoriquement égal au vecteur $[X_N]$ obtenu au bout de N itérations de Richardson (XII - 2) pour un même vecteur initial $[X_0]$.

$$[[X_N] - [X]] = [P_N(A)] [[X_0] - [X]]$$

avec $P_N(x) = \frac{T_N\left(\frac{-2x}{\alpha - \beta} + \frac{\alpha + \beta}{\alpha - \beta}\right)}{T_N\left(\frac{\alpha + \beta}{\alpha - \beta}\right)}$

Nous devons alors différencier la méthode de Stiefel Frank de celle de Richardson.

En s'appuyant sur le paragraphe précédent (1.2.3.1), on peut démontrer que l'itéré $[X_k]$ introduit par (XVIII-2) vaut

$$[X_k] = [P_k(A)] [X_0] - [Q_{k-1}(A)] [B] \quad k > 0 \quad (XIX - 2)$$

Or cet itéré peut être considéré comme élément d'une suite vectorielle dont le terme est justement (XIX - 2), suite convergente vers $[A^{-1}] [B]$

* cette suite est convergente

$$[V_k] = [[X_k] - [X_{k-1}]] - [[P_k(A)] - [P_{k-1}(A)]] [X_0] - [[Q_{k-1}(A)] - [Q_{k-2}(A)]] [B].$$

$$= [[P_k(A)] - [P_{k-1}(A)]] [[X_0] - [A^{-1}] [B]]$$

si $\omega = \frac{-2x + \alpha + \beta}{\alpha - \beta}$ et $h_k = \frac{1}{T_k\left(\frac{\alpha + \beta}{\alpha - \beta}\right)}$

$$\|V_k\| \leq \max_{x \in [\beta, \alpha]} |h_k T_k(\omega) - h_{k-1} T_{k-1}(\omega)| \| [X_0] - [A^{-1}] [B] \| \|\Omega\| \|\Omega^{-1}\|$$

$$\|V_k\| \leq 2 h_{k-1} K \quad \text{avec} \quad K = \| [X_0] - [A^{-1}] [B] \| \|\Omega\| \|\Omega^{-1}\|$$

La série $\sum V_k$ converge donc, car majorée par la série convergente de terme général positif $(2 h_{k-1} K)$

* Cette suite converge vers $[X] = [A^{-1}] [B]$

$$\text{En effet par construction } [[X_k] - [X]] = [P_k(A)] [[X_0] - [X]]$$

$$\text{donc} \quad \lim_{k \rightarrow \infty} [[X_k] - [X]] = [0]$$

La méthode de Stiefel-Frank permet donc d'atteindre la solution du système $[A] [X] = [B]$, comme limite d'une suite convergente, chacun des itérés $[X_k]$ se déterminant en fonction des deux itérés précédents $[X_{k-1}]$ et $[X_{k-2}]$ à l'aide de (XVIII - 2).

Ceci élimine le phénomène d'instabilité qui pouvait se produire, dans la méthode de Richardson, dont les itérés ne pouvaient pas être considérés comme termes d'une suite convergente.

L'influence des erreurs de chute, sera donc presque nulle. Ces erreurs n'empêcheront jamais la convergence de la suite, elles pourront néanmoins la ralentir, comme nous le verrons ultérieurement.

Variante de la Méthode de Stiefel-Frank

J. Albrecht (réf. 8) propose, par l'intermédiaire d'itérations du type (XVIII - 2), d'atteindre la solution du système $[A] [X] = [B]$ en prenant pour vecteur initial le vecteur nul.

$$[X_0] = [0]$$

$$[X_1] = \frac{2}{\alpha + \beta} [B]$$

La solution approchée sera encore $[X^*] = \lim_{k \rightarrow \infty} [X_k]$

$$\text{Mais } [X_k] = [P_k(A)] [X_0] - [Q_{k-1}(A)] [B]$$

$$[X_k] = -[Q_{k-1}(A)] [B] \quad \text{car} \quad [X_0] = [0]$$

$[-Q_{k-1}(A)]$ est donc l'approximation de $[A^{-1}]$, par un polynôme matriciel de degré $k-1$, au sens de Tchebycheff.

(le polynôme $-Q_{k-1}(x)$ est le polynôme qui approche $\frac{1}{x}$ sur $[\beta, \alpha]$, au sens de Tchebycheff).

2°) Comparaison entre la Méthode étudiée selon la technique R.E.C. et la méthode de Stiefel-Frank

La méthode proposée par Stiefel et W.L. Frank est vis à vis de la technique R.E.C. ce qu'était la méthode de Richardson en regard de la technique R.A.C.

Les itérations qu'elle utilise, ne jouent dans la technique R.E.C. qu'un rôle de sous itérations intermédiaires permettant le passage d'un itère de la suite $[X_k] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B]$ à l'itére suivant.

En terme de limite de suite vectorielle :

Méthode de Stiefel-Frank

$$[X^*] = \lim_{k \rightarrow \infty} [X_k] = \lim_{k \rightarrow \infty} [[P_k(A)] [X_0] - [Q_{k-1}(A)] [B]]$$

les $[X_k]$ se calculant par (XVIII - 2)

Méthode proposée (technique R.E.C.)

$$[X^*] = \lim_{k \rightarrow \infty} [X_k] = \lim_{k \rightarrow \infty} [[P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B]]$$

le passage de $[X_k]$ à $[X_{k-1}]$ se faisant par l'intermédiaire de p sous itérations définies par (XV - 2)

1. 2. 3. 3. Méthode semi-itérative de Tchebycheff

1°) Description

Considérons l'itération du type

$$[X^{(m+1)}] = [M] [X^{(m)}] + [G] \quad m \geq 0 \quad (XX - 2)$$

$$[X^{(0)}] \text{ donné}$$

où M est une matrice de rang n, vérifiant

$$\rho(M) = \max_{i=1, n} |u_i| < 1 \quad (u_i \text{ étant les valeurs propres de } M)$$

le procédé itératif (XX - 2) converge vers le vecteur $[X]$ solution de

$$\left. \begin{aligned} [X] &= [M] [X] + [G] \\ \text{ou} \quad [I] - [M] [X] &= [G] \end{aligned} \right\} \quad (XXI - 2)$$

Si $[\varepsilon^{(m)}] = [X^{(m)} - X]$ on vérifie que $[\varepsilon^{(m)}] = [M^m] [\varepsilon^{(0)}]$

$$\boxed{[X^{(m)} - X] = [\varepsilon^{(m)}] = [M^m] [\varepsilon^{(0)}] \quad m \geq 0.} \quad (XXII - 2)$$

Une méthode semi-itérative consiste à rechercher la solution comme combinaison linéaire des itérés (XX - 2).

$$[Y^{(m)}] = \sum_{j=0}^m \gamma_j(m) [X^{(j)}] \quad m \geq 0 \quad (\text{XXIII} - 2)$$

les variables $\gamma_j(m)$ étant choisies de façon à rendre

$$[E^{(m)}] = [Y^{(m)} - X] \quad \text{minimum}$$

Si $[X^{(m)}]$ est solution de (XXI - 2), $[Y^{(m)}]$ doit le rester

$$\Rightarrow \sum_{j=0}^m \gamma_j(m) = 1 \quad (\text{XXIV} - 2)$$

$$\begin{aligned} [E^{(m)}] &= [Y^{(m)} - X] = \sum_{j=0}^m (\gamma_j(m) [X^{(j)}]) - \left(\sum_{j=0}^m \gamma_j(m) \right) [X] \\ &= \sum_{j=0}^m \gamma_j(m) [X^{(j)} - X] = \sum_{j=0}^m \gamma_j(m) [\epsilon^{(j)}] \end{aligned}$$

$$[E^{(m)}] = \left(\sum_{j=0}^m \gamma_j(m) [M]^j \right) [\epsilon^{(0)}]$$

$$[E^{(m)}] = [P_m(M)] [\epsilon^{(0)}]$$

$$\text{si} \quad [P_m] = \sum_{j=0}^m \gamma_j(m) [M]^j \quad (\text{XXV} - 2)$$

$$\Rightarrow \|E^{(m)}\| \leq \|P_m(M)\| \|\epsilon^{(0)}\|$$

Si M est diagonalisable, $\|P_m(M)\| = \rho(P_m(M))$

(et si les matrices de transformation sont orthogonales).

$$= \max_{1 \leq i \leq n} |P_m(u_i)|$$

si les valeurs propres de $[M]$ sont localisées dans un intervalle $[\alpha, b]$ avec

$$-1 < \alpha \leq u_i \leq b < 1$$

Le polynôme choisi sera celui pour lequel $\max_{x \in [\alpha, b]} |P_m(x)|$ est minimum

$$P_m(x) = \frac{T_m\left(\frac{2x - (b+\alpha)}{b-\alpha}\right)}{T_m\left(\frac{2 - (b+\alpha)}{b-\alpha}\right)}$$

La méthode semi-itérative de Tchebycheff, (voir Varga (réf. 3)), considère le polynôme $p_m(x)$ minimum sur $[-\rho, +\rho]$ avec $p_m(1) = 1$

$$\text{et } \rho = \rho(M) = \text{Max } \{ |a_i| |b| \}$$

Le polynôme devient

$$p_m(x) = \frac{T_m\left(\frac{x}{\rho}\right)}{T_m\left(\frac{1}{\rho}\right)} \quad (\text{XXVI} - 2)$$

La relation de récurrence (R3) des polynômes de Tchebycheff, permet de construire directement les $[Y^{(m)}]$ de (XXIII - 2), sans préalablement calculer les $[X^{(m)}]$ de (XX - 2).

L'itération de la méthode semi-itérative de Tchebycheff est alors

$$\begin{aligned} [Y^{(m+1)}] &= \omega_{m+1} [[M] [Y^{(m)}] + [G] - [Y^{(m-1)}]] + [Y^{(m-1)}] \quad m \geq 1 \\ [Y^{(0)}] &= [X^{(0)}] \\ [Y^{(1)}] &= [M] [X^{(0)}] + [G] \quad (\text{XXVII} - 2) \\ \text{avec } \omega_1 &= 1. \\ \omega_{m+1} &= \frac{2 T_m(1/\rho)}{\rho T_{m+1}(1/\rho)} = 1 + \frac{T_{m-1}(1/\rho)}{T_{m+1}(1/\rho)} \quad m \geq 1 \end{aligned}$$

Le vecteur itère $[Y^{(m)}]$ est tel que

$$[E^{(m)}] = [[Y^{(m)}] - [X]] = [P_m(M)] \cdot [X^{(0)} - X]$$

$$\| [Y^{(m)}] - [X] \| < \frac{1}{T_m(1/\rho)} \| [X^{(0)}] - [X] \| \quad (\text{XXVII} - 2 \text{ bis})$$

2°) Remarques

A) La méthode objet de notre étude permet de résoudre un système linéaire $[A] [X] = [B]$,

la seule contrainte imposée est que $[A]$ soit diagonalisable et ait toutes ses valeurs propres réelles et de même signe.

La méthode semi-itérative de Tchebycheff ne permet de résoudre directement qu'un système $[X] = [M] [X] + [B]$ ou $[I-M] [X] = [A] [X] = [B]$ pour lequel la matrice $[M]$ est diagonalisable et $\rho(M) < 1$.

Elle ne permettra donc de résoudre un système $[A] [X] = [B]$ que si les valeurs propres de $[A] \in]0, 2[$.

- m itérations du type (XXVII - 2) avec $[M] = [I - A]$ donneraient comme solution approchée, $[Y^{(m)}]$ défini par

$$\| [Y^{(m)}] - [X] \| \leq \frac{1}{T_m \left(\frac{1}{\rho} \right)} \| [X^0] - [X] \|$$

$$\text{Si } \lambda_i \in]0, 2[\quad \text{avec } 0 < \beta < \dots \lambda_i \leq \dots < \alpha < 2.$$

$$1 - \lambda_i \in [1 - \alpha, 1 - \beta]$$

$$\text{Donc } \rho = \rho(M) = \max (|1 - \alpha|, |1 - \beta|)$$

- m itérations secondaires de Richardson, (du type XII - 2) ou secondaires de la technique de Stiefel-Frank (XVIII - 2), correspondant à 1 itération principale de la méthode étudiée, donneront un vecteur $[X_m]$ défini par

$$\| [X_m] - [X] \| \leq \frac{1}{T_m \left(\frac{\alpha + \beta}{\alpha - \beta} \right)} \| [X^0] - [X] \|^2$$

pour un même vecteur de départ $[X^0]$

Le vecteur $[X_m]$ ainsi obtenu sera plus proche de la solution que le vecteur $[Y^{(m)}]$ car on peut vérifier que $\frac{\alpha + \beta}{\alpha - \beta} > \frac{1}{\rho}$

B) On pourrait concevoir d'utiliser la méthode semi-itérative de Tchebycheff pour résoudre le système $[X] = [P_p(A)] [X] - [Q_{p-1}(A)] [(B)]$, que nous savons équivalent à $[A] [X] = [B]$. si $p(0) = 1$

$$\begin{aligned} p(x) &< 1 \text{ sur } [\beta, \alpha] \\ q(x) &= (p(x) - 1) / x \end{aligned}$$

$$\begin{cases} [M] = [P_p(A)] \\ [G] = -[Q_{p-1}(A)] [B] \end{cases}$$

L'itération (XVII - 2) donnerait alors

$$\begin{aligned} [Y^{(m+1)}] &= \omega_{m+1} [[P_p(A)] [Y^{(m)}] - [Q_{p-1}(A)] [B] - [Y^{(m-1)}]] + [Y^{(m-1)}] \\ \text{avec} \quad [Y^0] &= [X_0] \quad \quad \quad \text{(XXVIII - 2)} \\ [Y^1] &= [X_1] = [P_p(A)] [X_0] - [Q_{p-1}(A)] [B] \end{aligned}$$

m itérations (XVII - 2) donneront un vecteur $[Y^{(m)}] = \sum_{j=0}^m \gamma_j^{(m)} [X_j]$

$$\text{tel que } \|[Y^{(m)}] - [X]\| \leq \frac{1}{T_m \left(\frac{1}{\rho}\right)} \|[X_0] - [X]\|$$

$$\text{or } \rho = \rho(M) = \rho(P_p(A)) = \frac{1}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta}\right)}$$

$$\|[Y^{(m)}] - [X]\| \leq \frac{1}{T_m \left(T_p \left(\frac{\alpha + \beta}{\alpha - \beta}\right)\right)} \|[X_0] - [X]\|$$

$$\|[Y^{(m)}] - [X]\| \leq \frac{1}{T_{mp} \left(\frac{\alpha + \beta}{\alpha - \beta}\right)} \|[X_0] - [X]\| \quad \begin{array}{l} \text{en tenant compte de la} \\ \text{relation R2 des polynômes} \\ \text{de Tchebycheff (1.1.5.3.)} \end{array}$$

Le vecteur ainsi obtenu, est le même que celui qui aurait été obtenu en une itération principale de la méthode étudiée lorsque le degré du polynôme est m, p c'est-à-dire en mp itérations de Richardson (RAC), ou mp itérations de Stiefel-Frank (REC).

1. 3. PROBLEME DE L'OPTIMISATION

1. 3. 1. Position du problème

La méthode de suritération polynomiale, objet de notre étude, nécessite, quelle que soit la technique adoptée, N itérations principales (III - 1), chacune d'elles se décomposant en P "sous-itérations", soit du type (X - 2), soit du type (XVI - 2).

La résolution numérique d'un système linéaire $[A][X] = [B]$ de n équations à n inconnues nécessite donc le calcul de $P \times N$ produits du type $[A][X_i]$.

Optimiser la résolution consiste, théoriquement, à *déterminer au mieux les deux entiers N et P* de façon à rendre le vecteur final $[X_N] = [X^*]$ (dans la suite $[X^*]$ représentera toujours la solution approchée du système)

- soit aussi proche que possible de la solution exacte
: *optimisation en qualité*
- soit en un temps minimum
: *optimisation en rapidité de calcul*
- soit en nécessitant une occupation de mémoire qui soit minimum
: *optimisation en occupation de mémoire*

Si les calculs s'effectuaient suivant les prévisions théoriques, la méthode à adopter en regard des deux premières optimisations serait, soit la méthode de Richardson, soit la méthode de Stiefel-Frank, ce qui imposerait pour N la valeur 1.

La 3ème optimisation donnerait la préférence à la méthode de Richardson, qui exige du fait des itérations qu'elle propose, la réservation d'un vecteur en moins.

Les erreurs de troncature introduites au cours des calculs, ont pour effet, par leur cumul, que la solution $[X_N] = [X^*]$ effectivement obtenue n'est jamais égale à celle qui devrait être théoriquement obtenue. Elle peut parfois en être très éloignée.

Quitte à y perdre en rapidité et en occupation de mémoire, nous préférons avant tout préciser la signification des résultats obtenus en contrôlant le déroulement du processus itératif.

Les formules de majoration d'erreur, introduites dans le premier chapitre (1.1.4.), seront utilisées à cet effet.

Ce contrôle d'erreur, pourra néanmoins n'être pas fait, lorsque le système est "facilement résoluble". Le critère "facilement", purement empirique, sera fonction du rapport des valeurs propres extrêmes $y = \frac{\alpha}{\beta}$.

Plus γ est élevé, plus la résolution est longue, et exige une quantité de calculs importante, et donc plus les erreurs de troncature risquent d'influer sur le résultat. Un contrôle des erreurs sera alors nécessaire.

Remarque : Dans ce chapitre, nous n'étudierons que les systèmes pour lesquels

$$\|P_p(A)\| = \rho(P_p(A)) = \max_i |p_p(\lambda_i)|.$$

1. 3. 2. Détermination du degré P du polynôme $[P_p(A)]$

1. 3. 2. 1. Détermination mathématique

Si nous voulons atteindre en N "itérations principales", une solution $[X^*] = [X_N]$ telle que

$$\|[X] - [X^*]\| \leq \varepsilon \|[X] - [X_0]\|$$

, il est possible, en supposant connus α et β , donc $\gamma = \frac{\alpha}{\beta}$, de calculer à une unité près, le degré P à affecter au polynôme $[P_p(A)]$.

$$\text{En effet } \varepsilon = \|P_p(A)\|^N = h^N(p, \alpha / \beta) = h^N$$

$$h = \varepsilon^{1/N}$$

$$h = \frac{1}{T_p\left(\frac{\alpha + \beta}{\alpha - \beta}\right)} = \frac{1}{\text{chp } \omega}$$

$$\text{avec } \omega = \text{Argt ch}\left(\frac{\alpha + \beta}{\alpha - \beta}\right) = \text{Argt ch}\left(\frac{\gamma + 1}{\gamma - 1}\right)$$

$$\text{Argt ch } u = \text{Log}(u + \sqrt{u^2 - 1})$$

$$\omega = \text{Log}\left(\frac{\sqrt{\gamma + 1}}{\sqrt{\gamma - 1}}\right)$$

$$\frac{1}{\text{chp } \omega} = \varepsilon^{1/N} \quad p = \frac{1}{\omega} \text{ Argt ch } (\varepsilon^{-1/N})$$

$$= \frac{\text{Log}(\varepsilon^{-1/N} + \sqrt{\varepsilon^{-2/N} - 1})}{\text{Log}\left(\frac{\sqrt{\gamma + 1}}{\sqrt{\gamma - 1}}\right)}$$

P devant être entier

$$P = \text{ENTIER} \left[\frac{\text{Log} (\varepsilon^{-1/N} + \sqrt{\varepsilon^{-2/N} - 1})}{\text{Log} \left(\frac{\sqrt{Y} + 1}{\sqrt{Y} - 1} \right)} \right] \quad (I - 3)$$

1. 3. 2. 2. Détermination graphique

La résolution d'un grand système linéaire, nécessitera le plus souvent une étude préalable : calcul des valeurs propres α et β , puis choix de la technique à adopter.

Le calcul de P par (I.3), n'étant pas aisé, une détermination graphique du degré peut être plus commode.

A cet effet des réseaux de courbe $h(P, Y) = K = \text{cste}$, ont été tracés pour différentes valeurs de K.

Ces réseaux ont été groupés sur trois graphiques

- le premier donnant les courbes $h=K$ pour $\left| \begin{array}{l} Y \in [1, 100] \\ P \leq 80 \end{array} \right.$
- le deuxième pour $\left| \begin{array}{l} Y \in [100, 2500] \\ P \leq 400 \end{array} \right.$
- le troisième pour $\left| \begin{array}{l} Y \in [2500, 25000] \\ P \leq 1000 \end{array} \right.$

Une simple lecture fournira les valeurs N et P correspondant à $\varepsilon = h^N(P, Y)$

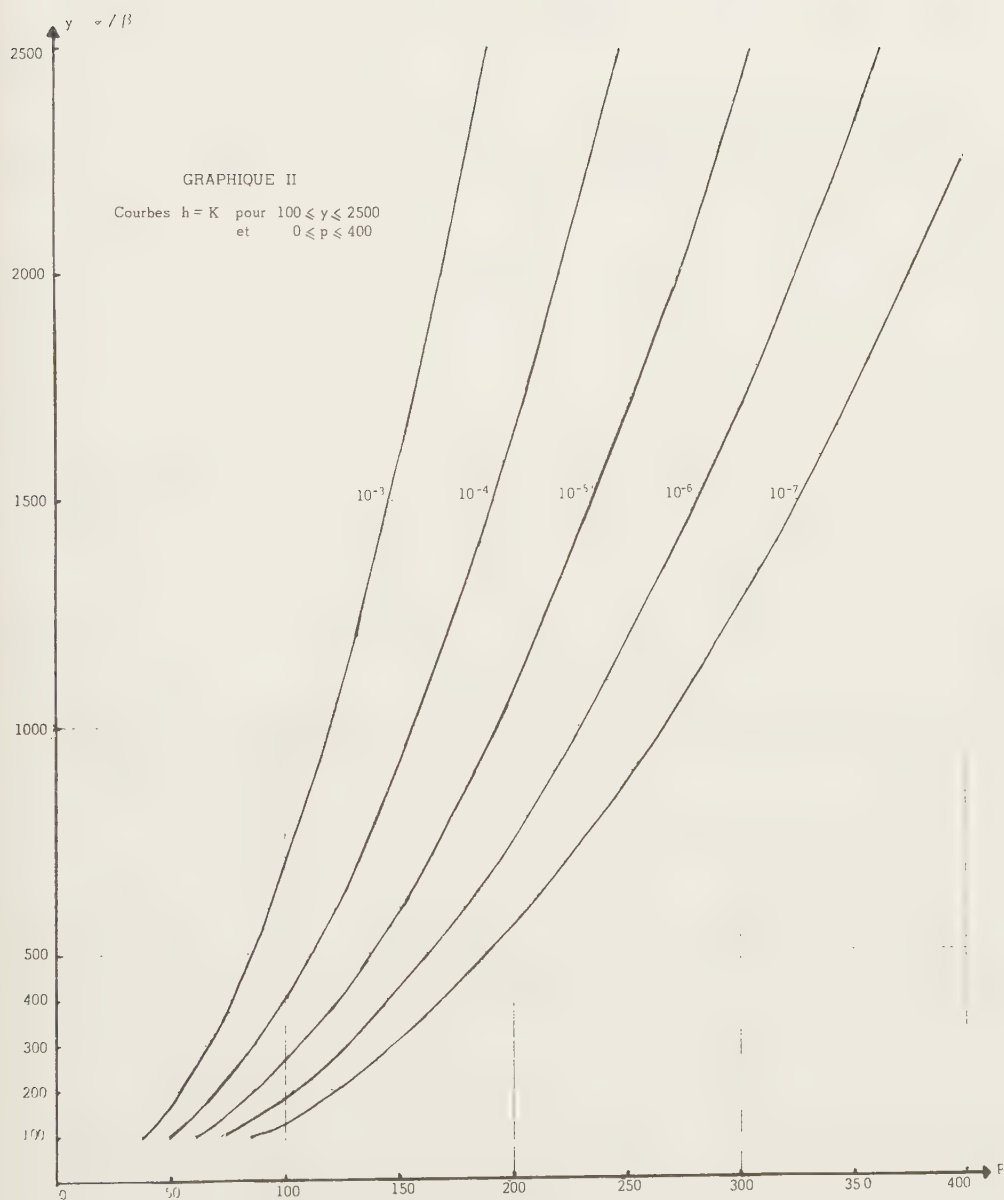
Exemples : 1) $Y = 2000$

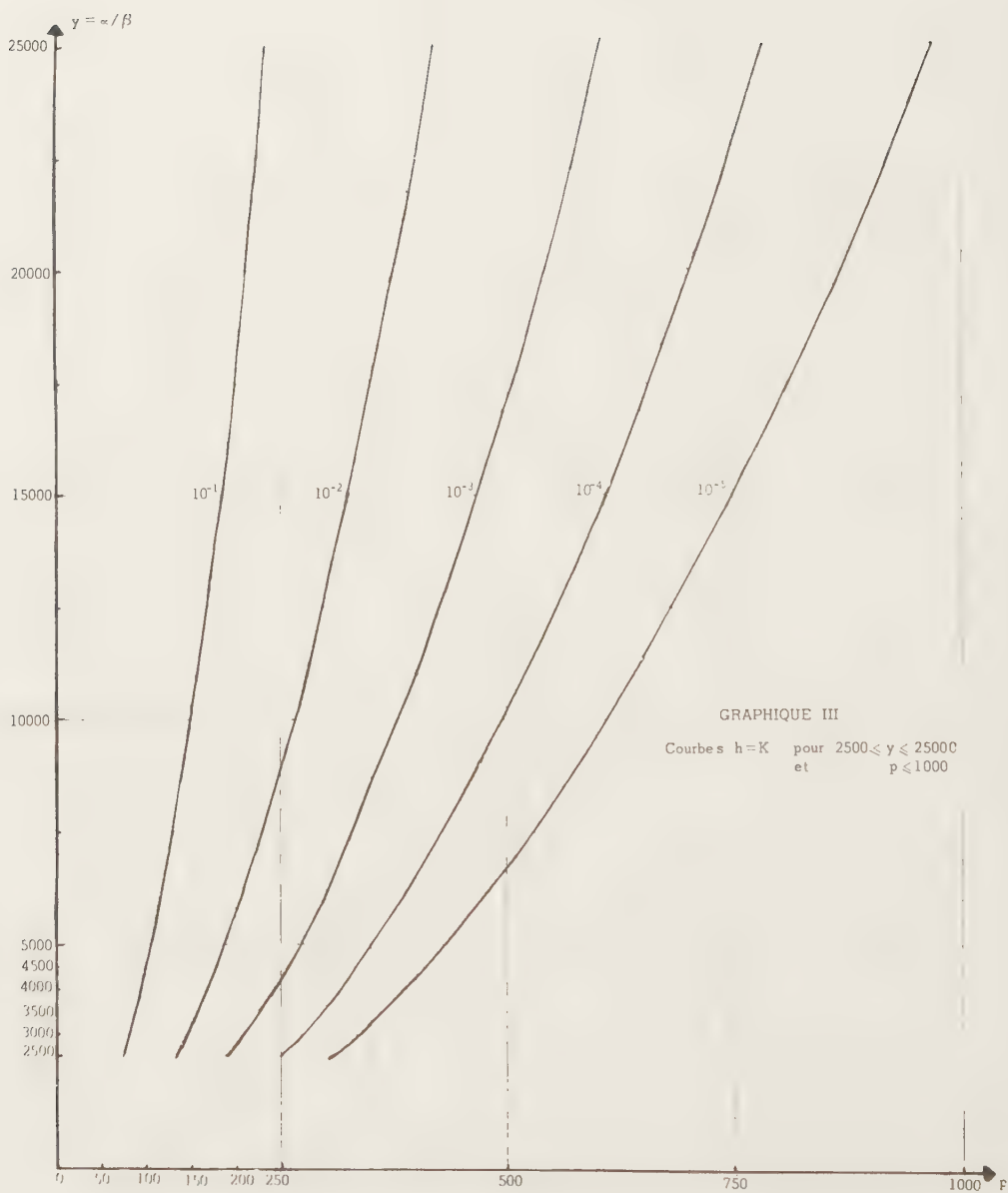
$\varepsilon = 10^{-6}$ peut être obtenu avec $\left| \begin{array}{l} N = 1 \text{ et } P = 325 \\ N = 2 \text{ et } P = 170 \\ N = 3 \text{ et } P = 120 \end{array} \right.$

2) $Y = 20000$

$\varepsilon = 10^{-4}$ peut être obtenu avec $\left| \begin{array}{l} N = 1 \text{ et } P = 700 \\ N = 2 \text{ et } P = 375 \\ N = 4 \text{ et } P = 215 \end{array} \right.$







1. 3. 3. Influence des erreurs d'arrondi sur un procédé itératif.

1. 3. 3. 1. Résolution d'une équation par technique itérative.

La résolution d'une équation par procédé numérique, fait toujours intervenir un système itératif qui peut être mis sous la forme $U_{N+1} = MU_N$ la convergence nécessitant $|M| < 1$.

Si la convergence est franche U_{N+1} tendra très rapidement vers zéro.

Le plus souvent des erreurs de chute viennent altérer le processus itératif. Leur présence peut être représentée à chaque itération par un terme d'erreur ϵ_i , aléatoire.

$$\left| \begin{array}{l} U_1 = MU_0 + \epsilon_1 \\ U_2 = MU_1 + \epsilon_2 \\ \text{-----} \\ U_N = MU_{N-1} + \epsilon_N \\ U_{N+1} = MU_N + \epsilon_{N+1} \end{array} \right.$$

Les ϵ_i , par cumul, se répercutent sur U_{N+1} , qui devient

$$U_{N+1} = M^{N+1} U_0 + \epsilon_{N+1} + \epsilon_N M + \dots + \epsilon_1 M^N$$

$$U_{N+1} = M^{N+1} U_0 + R$$

$$\text{avec } R = \epsilon_{N+1} + \epsilon_N M + \dots + \epsilon_1 M^N$$

Le comportement de R est aléatoire, mais si nous supposons tous les ϵ_i majorés en valeur absolue par ϵ

$$|\epsilon_i| \leq \epsilon \quad (\epsilon > 0)$$

$$\text{alors } |R| \leq \epsilon (1 + |M| + |M|^2 + \dots + |M|^N)$$

$$\text{soit } |R| \leq \epsilon \left(\frac{1 - |M|^{N+1}}{1 - |M|} \right)$$

Lorsque $N \rightarrow \infty$, R reste majoré par $R_0 = \frac{\epsilon}{1 - |M|}$

ce qui se traduit par $|U_{N+1} - M^{N+1} U_0| \leq R_0$

au lieu de tendre vers 0, U_{N+1} , pour N infini, prend un comportement aléatoire et oscille entre R_0 et $-R_0$.

Ce phénomène limite la précision à atteindre par cette technique itérative qui sera d'autant moins bonne que M sera proche de 1.

Au test d'arrêt habituel $|U_N| \leq \text{EPS}$, il faudra joindre un autre test permettant de déceler si le "seuil de précision" R_0 est atteint.

$$\text{Ce test pourra être } \left| \frac{U_{N+1}}{U_N} \right| \geq 1$$

1. 3. 3. 2. Résolution d'un système linéaire par méthode itérative.

Une étude analogue peut être faite dans le cas d'une résolution de n équations à n inconnues. Appliquons la à la méthode de résolution qui nous préoccupe.

Si $[X_i] = [P_p(A)] [X_{i-1}] - [Q_{p-1}(A)] [B]$, la convergence peut être étudiée par la relation

$$[U_{N+1}] = [P_p(A)] [U_N] \text{ avec } [U_N] = [X_N] - [X_{N-1}]$$

Les itérations effectives définissant $[U_{N+1}]$ pourront être écrites

$$\begin{cases} [U_1] = [P_p(A)] [U_0] + [F_1] \\ [U_2] = [P_p(A)] [U_1] + [F_2] \\ \dots\dots\dots \\ [U_N] = [P_p(A)] [U_{N-1}] + [F_N] \\ [U_{N+1}] = [P_p(A)] [U_N] + [F_{N+1}] \end{cases}$$

Les vecteurs "erreurs" $[F_i]$, se répercuteront de proche en proche, et $[U_{N+1}]$, au lieu de valoir $[P_p(A)]^{N+1} [U_0]$, vaudra en fait

$$[U_{N+1}] = [P_p(A)]^{N+1} [U_0] + [F_{N+1}] + [P_p(A)] [F_N] + \dots + [P_p(A)]^N [F_1]$$

nous pouvons supposer les $[F_i]$ majorés en norme par une quantité E .

$$\forall i \quad \|F_i\| \leq E$$

$$\|[U_{N+1}] - [P_p(A)]^{N+1} [U_0]\| \leq \frac{E(1-h^{N+1})}{(1-h)} \quad \text{où } h = \rho(P_p(A))$$

Lorsque $N \rightarrow \infty$, $[U_{N+1}]$ reste tel que

$$\|[U_{N+1}] - [P_p(A)]^{N+1} [U_0]\| \leq \frac{E}{1-h}$$

Le vecteur $[U_{N+1}]$, au lieu de tendre vers le vecteur nul, peut prendre un comportement aléatoire se manifestant par une oscillation de chacune de ses composantes autour de zéro, entre ϵ_i et $-\epsilon_i$.

$$(\text{avec } \sqrt{\sum \epsilon_i^2} = R = \frac{E}{1-h})$$

Pour un système linéaire, bien défini, c'est-à-dire défini par le rapport $y = \frac{\alpha}{\beta}$, la précision finale obtenue dépendra de la valeur prise par $R = \frac{E}{1-h}$.

$$1^{\circ}) \text{ Etude de la fonction } R = \frac{E}{1-h}$$

Nous connaissons parfaitement le comportement de h , donc celui de $\frac{1}{1-h}$, qui ne sera, y étant fixé, que fonction du degré p .

Il est naturel de penser, puisqu'une itération principale, se compose de p sous itérations, que chacun des $[F_i]$, donc E , est fonction de p . Il est aussi logique de supposer que E est fonction croissante de p .

E dépendra de plus de la technique adoptée (RAC et REC). Si $y = \frac{\alpha}{\beta}$ est faible, $h(p)$ décroît très rapidement lorsque p augmente, et la convergence exige un nombre de sous itérations peu élevé. Nous pouvons supposer que cette convergence est franche et rapide.

Il n'en est certainement pas de même lorsque y atteint de très grandes valeurs. Un degré p très élevé est en effet nécessaire pour que $h(p)$ diffère sensiblement de 1, donc pour que $\frac{1}{1-h}$ ne soit pas très grand. Mais alors $E(p)$ risque d'être important, ainsi que $R(p)$.

Le problème consistera alors à rechercher le degré p rendant minimum la fonction $\frac{E(p)}{1-h} = R(p)$.

Deux réseaux de courbe $R(p)$ pour des y différents sont représentés sur les graphiques IV et V. Ils correspondent chacun à deux fonctions $E(p)$ croissantes :

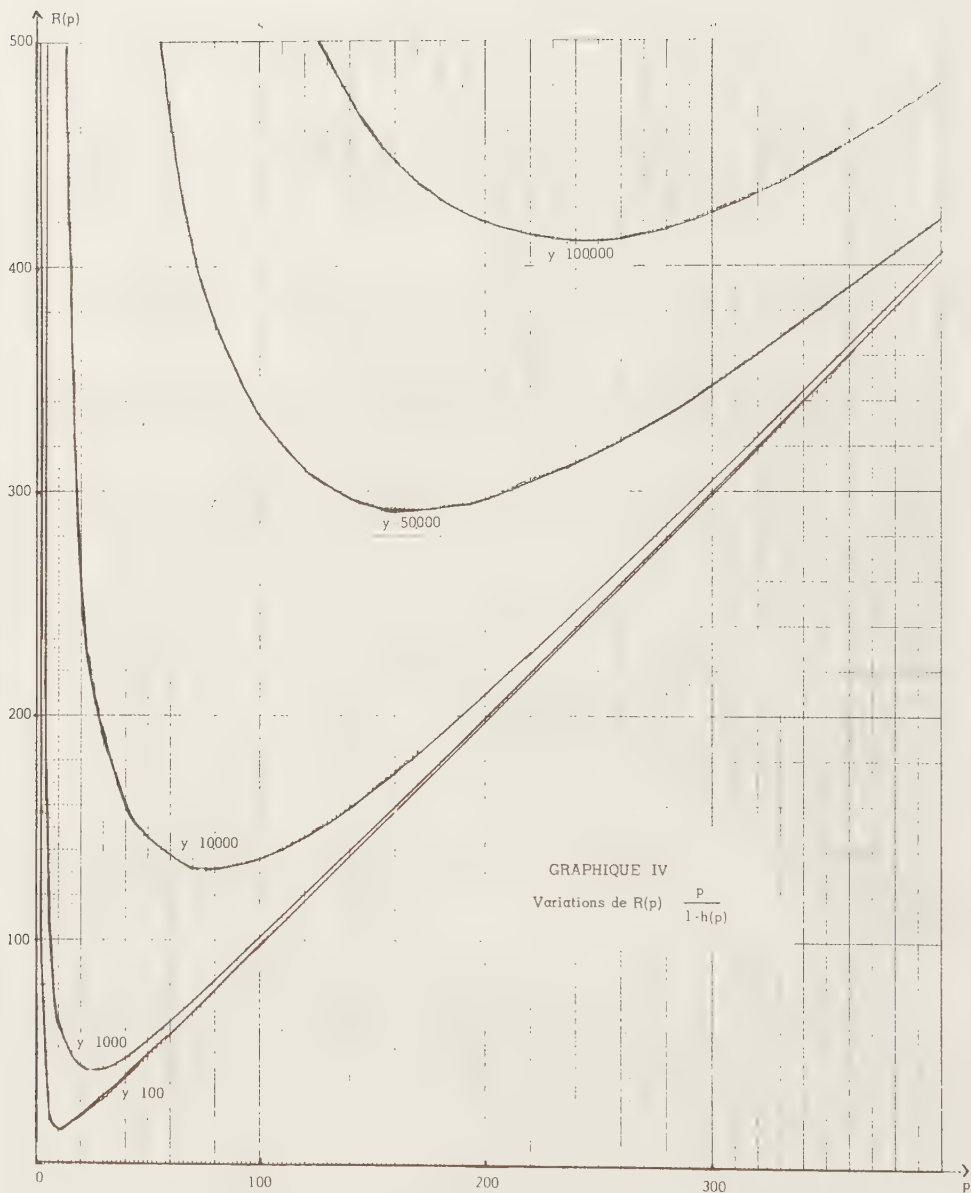
$$E(p) = p$$

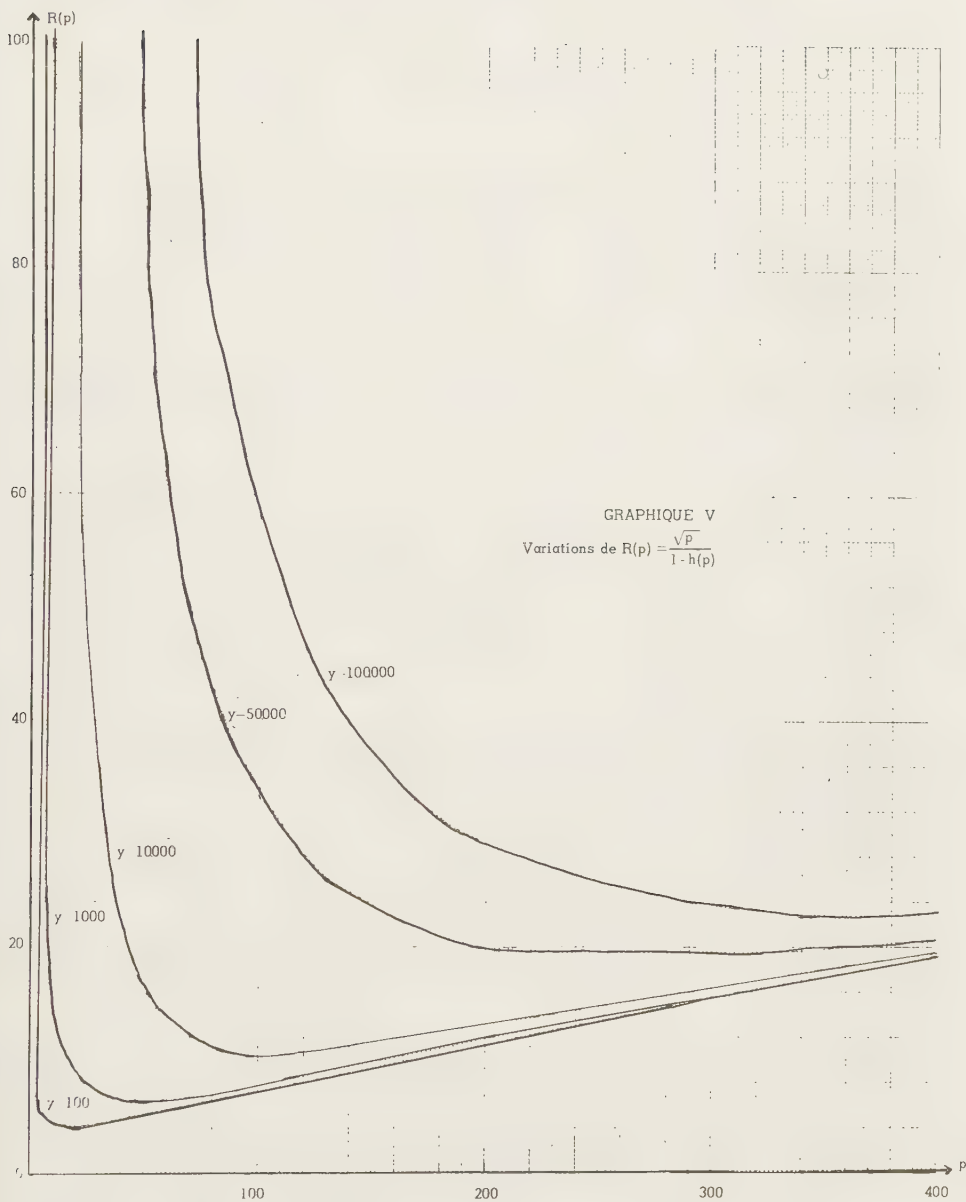
$$\text{et } E(p) = \sqrt{p}.$$

Ces courbes, qui ont un comportement identique, montrent bien qu'un minimum est atteint pour un p^* . Elles indiquent de plus que cette valeur minimum p^* augmente lorsque y augmente, et parallèlement, $R(p^*)$,

Plus y est élevé, plus la précision finale qu'il est possible d'atteindre est faible.

Cette étude est assez hypothétique, mais nous ne pouvons dans l'état de nos connaissances actuelles sur les erreurs d'arrondi et leur influence, qu'émettre des hypothèses.





1. 3. 4. Contrôle des erreurs d'arrondi

L'utilisation des formules de majoration d'erreurs, introduites au paragraphe 1. 1. 4. va nous permettre de vérifier si le processus itératif s'effectue de façon conforme à la théorie.

$$\text{Si } [X_i] = [P_p(A)] [X_{i-1}] - [Q_{p-1}(A)] [B]$$

$$\text{et } [U_i] = [X_i] - [X_{i-1}]$$

, le vecteur $[X_{i+1}]$ doit être tel que

$$\frac{\|X_{i+1} - X_i\|}{\|X_i - X_{i-1}\|} = \frac{\|U_{i+1}\|}{\|U_i\|} \leq h \quad \text{avec } h = \frac{1}{T_p\left(\frac{\alpha+\beta}{\alpha-\beta}\right)} = \|P_p(A)\|$$

$$(K = \|\Omega\| \cdot \|\Omega^{-1}\| = 1).$$

à chaque itération, l'influence des erreurs d'arrondi, pourra être étudiée grâce à la comparaison du rapport $\frac{\|U_{i+1}\|}{\|U_i\|}$ avec h.

Si ces deux grandeurs vérifient la relation définie ci-dessus ou du moins sont comparables, nous pourrions considérer que les erreurs d'arrondi n'altèrent pas les itérations. Si elles ne le sont plus, le processus devra être stoppé et si l'arrêt s'effectue avec $\frac{\|U_N\|}{\|U_{N-1}\|} \gg 1$, la solution finale retenue sera $[X_{N-1}] = [X^*]$

s'il s'effectue avec $h \leq \frac{\|U_N\|}{\|U_{N-1}\|} \leq 1$, la solution retenue sera $[X_N] = [X^*]$.

La relation (VIII - 1) permettra alors de préciser la signification du résultat retenu, (au sens absolu), vis à vis de la solution exacte, et

$$\begin{aligned} \|X - X^*\| &\leq \|U_{N-1}\| \cdot \frac{1}{1-h} \quad \text{si } [X^*] = [X_{N-1}] \\ \|X - X^*\| &\leq \|U_N\| \cdot \frac{1}{1-h} \quad \text{si } [X^*] = [X_N] \end{aligned} \quad (\text{II} - 3)$$

Une *précision relative* du résultat pourra être aussi donnée, dans le cas où le vecteur de départ $[X_0]$ est nul.

En effet nous savons que

$$\begin{aligned} \|X - X_k\| &\leq h \|X - X_{k-1}\| \\ \|X - X_k\| &\leq h^k \|X - X_0\| \\ \frac{\|X - X_k\|}{\|X\|} &\leq h^k. \end{aligned}$$

D'où les valeurs donnant les précisions relatives :

$$\begin{array}{l} \frac{\|X - X^*\|}{\|X\|} \leq h^{N-1} \quad \text{si } [X^*] = [X_{N-1}] \\ \frac{\|X - X^*\|}{\|X\|} \leq h^N \quad \text{si } [X^*] = [X_N] \end{array} \quad (\text{III} - 3)$$

1. 3. 5. Résolution d'un système linéaire "franc".

1. 3. 5. 1. Définition d'un système franc.

Nous définissons un système "franc", comme étant un système dont la résolution est rapide et de qualité.

Nous sous-entendons ainsi, que les erreurs de chute jouent un rôle qui peut être considéré comme négligeable. Elles ne freinent ni ne faussent la résolution.

Aucune règle théorique ne nous permet de distinguer à priori un système linéaire "franc", d'un système "non franc", c'est-à-dire difficile à résoudre : Une telle distinction ne peut être qu'empirique.

Nous considérons comme "franc" tout système linéaire résoluble en une seule itération principale.

Le balayage systématique des racines, que nous avons introduit dans la technique R.A.C. nous a toujours permis d'itérer (III - 1) avec des polynômes de degré 240 = P.

Remarque : Ce degré 240 que nous citons, est le degré maximum qu'il a toujours été possible d'utiliser. Des degrés supérieurs (400) ont pu parfois être traités.

L'ordinateur utilisé au cours de ces essais, un ordinateur CAE 510 nécessite pour la représentation d'un nombre en virgule flottante, deux mots de 18 bits, ce qui correspond à des mots de 8 chiffres significatifs.

Des ordinateurs travaillant avec des nombres de 10 chiffres ou plus, permettraient vraisemblablement l'emploi de degré plus élevés. Cette limite que nous avons ici définie doit donc être considérée comme fonction de l'ordinateur utilisé.

Si nous consultons les réseaux de courbe $h = K$, nous constatons qu'une itération de degré 240 permet d'obtenir pour un système linéaire défini par

$$y = \frac{\alpha}{\beta} \approx 10^3, \text{ une solution } [X_1] = [X^*] \text{ telle que } \|X - X^*\| \leq 10^{-6} \|X - X_0\|$$

Si $[X_0] = [0]$, vecteur initial qui a toujours été adopté lorsque nous n'avions à priori aucune idée sur la solution, la relation ci-dessus peut s'écrire

$$\frac{\|X - X^*\|}{\|X\|} \leq 10^{-6}$$

La précision relative ainsi obtenue peut être considérée comme très bonne, car elle se traduit en général par six chiffres significatifs exacts sur chaque composante de $[X^*]$.

Nous considérerons donc cette valeur $\gamma = 10^3$, comme étant la frontière délimitant les systèmes "francs" des autres.

1. 3. 5. 2. Choix de la technique à adopter.

Si nous ne sommes pas limités par la capacité en mémoire centrale de l'ordinateur utilisé, la technique à adopter sera sans contestation possible la technique R. E. C. avec $N=1$, c'est-à-dire la méthode de Stiefel Frank.

Elle aura sur l'autre technique l'avantage d'obtenir la précision maximum.

Sur ordinateur de faible capacité en mémoire centrale, la technique R. A. C. avec $N=1$, donc la méthode de Richardson avec balayage des racines, sera préférée, car elle permet d'obtenir avec gain d'un vecteur en mémoire centrale, une solution de qualité équivalente.

1. 3. 5. 3. Contrôle des erreurs de troncature.

Les méthodes conseillées ci-dessus, ne permettent pas de contrôler le processus itératif, le calcul de la quantité $\frac{\|X_i - X_{i-1}\|}{\|X_{i-1} - X_{i-2}\|}$ nécessitant au moins deux itérations principales.

Si le vecteur de départ est $[X_0] = [0]$, nous pouvons contrôler les erreurs de chute, sans y perdre en rapidité de calcul, et en précision finale, par *utilisation d'une itération de la méthode semi-itérative de Tchebycheff*.

Ce contrôle nécessitant la mise en réserve de l'itéré précédent fera alors préférer l'utilisation de la technique R. A. C.

Supposons qu'une itération de degré 2. $P_1 = P$ soit théoriquement nécessaire pour obtenir un vecteur solution, de précision relative $\leq h_{2P_1}$.

Construisons deux itérations principales de degré P_1 ($N=2$)

$$[X_1] = [P_{P_1}(A)] [X_0] - [Q_{P_1-1}(A)] [B] = -[Q_{P_1-1}(A)] [B] \quad \text{car } [X_0] = [0]$$

$$[X_2] = [P_{P_1}(A)] [X_1] - [Q_{P_1-1}(A)] [B]$$

Si les erreurs d'arrondi n'ont joué qu'un rôle négligeable, c'est-à-dire si

$$\frac{\|X_2 - X_1\|}{\|X_1 - X_0\|} \cdot \frac{\|X_2 - X_1\|}{\|X_1\|} \leq h_{P_1}, \quad \text{le vecteur } [X_2] \text{ obtenu est vis à vis de}$$

$$[X] = [A^{-1}] [B] \text{ tel que } \|X_2 - X\| \leq h_{P_1}^2 \|X_0 - X\|$$

$$\frac{\|X_2 - X\|}{\|X\|} \leq h_{P_1}^2$$

Construisons alors le vecteur $[Y^{(2)}]$, défini par la relation (XXVIII - 2), deuxième itère, de la méthode semi-itérative de Tchebycheff.

$$\text{avec } \begin{cases} [M] = [P_{p_1}(A)] \\ [G] = [-Q_{p_1-1}(A)] [B] \end{cases}$$

(cf 1.2.3.3. 1er B)

$$[Y^{(2)}] = \omega_2 \{ [P_{p_1}(A)] [Y^{(1)}] - [Q_{p_1-1}(A)] [B] - [Y^{(0)}] \} + [Y^{(0)}]$$

$$\text{or } [Y^{(0)}] = [X_0] = [0]$$

$$\text{et } [Y^{(1)}] = [X_1] = [P_{p_1}(A)] [X_0] - [Q_{p_1-1}(A)] [B] - [-Q_{p_1-1}(A)] [B]$$

$$[Y^{(2)}] = \omega_2 [X_2]$$

$$\omega_2 = \frac{2 T_1 \left(\frac{1}{\rho} \right)}{\rho T_2 \left(\frac{1}{\rho} \right)}$$

$$\rho = \rho(M) = \rho(P_{p_1}(A)) = h_{p_1} \quad 2. \quad \frac{1}{h_{p_1}}$$

$$\begin{cases} T_1(u) = u \\ T_2(u) = 2u^2 - 1 \end{cases} \quad \omega_2 = \frac{1}{h_{p_1} \left(\frac{2}{h_{p_1}^2} - 1 \right)}$$

$$\boxed{[Y^{(2)}] = \frac{2}{2 - h_{p_1}^2} [X_2]} \quad (\text{III} - 3)$$

Nous savons (XXVII-2bis) que $[Y^{(2)}]$ est vis à vis de $[X] = [A^{-1}] [B]$

$$\text{tel que } \|Y^{(2)} - X\| \leq \frac{1}{T_2 \left(\frac{1}{\rho} \right)} \|X_0 - X\|$$

$$\frac{1}{T_2 \left(\frac{1}{\rho} \right)} = \frac{1}{T_2 \left(T_{p_1} \left(\frac{\alpha + \beta}{\alpha - \beta} \right) \right)} = \frac{1}{T_{2p_1} \left(\frac{\alpha + \beta}{\alpha - \beta} \right)} = h_{2p_1}$$

$$\boxed{\frac{\|Y^{(2)} - X\|}{\|X\|} \leq h_{2p_1}} \quad (\text{IV} - 3)$$

Donc, si $[X_0] = [0]$, deux itérations principales de degré p_1 , puis une simple multiplication du vecteur $[X_2]$ par une constante donnent dans un temps aussi rapide, et avec possibilité de contrôle d'erreur, un résultat aussi précis, que celui qui aurait été obtenu avec une itération de degré $2p_1$.

1. 3. 6. Résolution d'un système linéaire "non franc".

Rappelons que nous considérons comme tel, tout système tel que le rapport $y = \frac{\alpha}{\beta}$ soit $> 10^3$.

Un tel système exigera pendant sa résolution un contrôle des erreurs d'arrondi, ce contrôle s'effectuant par la comparaison, à chaque itération, des quantités h et $\frac{\|X_i - X_{i-1}\|}{\|X_{i-1} - X_{i-2}\|}$.

Pour un $\varepsilon = h^N(p,y)$ désiré, le nombre N d'itérations principales dépendra de la technique utilisée. Ce nombre sera de toute façon choisi ≥ 2 afin de permettre le contrôle des itérations.

Si la mémoire de l'ordinateur utilisé est limitée, la technique R.A.C. sera préférée. Le nombre N devra être alors choisi, à l'aide des courbes $h = K$, de façon que P n'excède pas 240 environ (voir la remarque par 1.3.5.1.).

Si la mémoire le permet, la technique R.E.C., autorisant des degrés P supérieurs, rendra la résolution un peu plus rapide. Le degré P ne devra néanmoins pas être très élevé, car la convergence de la technique R.E.C., si elle a lieu pour tout p , risque d'être parfois freinée à cause des erreurs de chute. (Un exemple sur des matrices test en témoignera).

De toutes façons, que l'une ou l'autre des deux techniques soit adoptée, si $[X_0] = [0]$, une itération de la méthode semi-itérative de Tchebycheff, pourra être utilisée comme en (1.3.5.3.) pour activer la convergence à la deuxième itération principale.

Mais alors, le test suivant, de contrôle de convergence, devra porter, non pas sur $\frac{\|X_3 - X_2\|}{\|X_2 - X_1\|} \leq h$, mais sur $\frac{\|X_3 - Y^{(2)}\|}{\|X_2 - X_1\|} \leq h$.

1. 4. Résolution de systèmes test

1. 4. 1. Matrice tridiagonale

La matrice du premier système étudié est une matrice tridiagonale, c'est-à-dire du type Jacobi. Cette matrice est bien connue. Elle ne diffère de la matrice "différence seconde" que par son premier élément (réf. 7, réf. 10).

$$[A_1] = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & & \\ 0 & -1 & 2 & -1 & \\ 0 & & \ddots & \ddots & \\ 0 & & & -1 & 2 \end{bmatrix}$$

Inverse de A_1

L'inverse de $[A_1]$ est très simple, ses éléments étant tous entiers et compris entre 1 et n , si n est sa dimension.

$$[A_1^{-1}] = \begin{bmatrix} n & n-1 & n-2 & \cdots & 2 & 1 \\ n-1 & n-1 & n-2 & \cdots & 2 & 1 \\ n-2 & n-2 & n-2 & & 2 & 1 \\ n-3 & n-3 & n-3 & & 2 & 1 \\ \vdots & \vdots & & & 2 & 1 \\ \vdots & \vdots & & & 2 & 1 \\ 2 & \cdots & \cdots & \cdots & 2 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \quad \left\{ \begin{array}{l} \alpha_{ij} = n+1-i \\ \quad \text{si } i \leq j \\ \alpha_{ji} = \alpha_{ij} \end{array} \right.$$

Valeurs propres de A_1

Les valeurs propres sont toutes positives et

$$\lambda_i = 4 \sin^2 \left(\frac{(2i-1)\pi}{2(2n+1)} \right) \quad (i=1, \dots, n)$$

$$\Rightarrow \left\{ \begin{array}{l} \alpha = 4 \sin^2 \left(\frac{(2n-1)\pi}{2(2n+1)} \right) \\ \beta = 4 \sin^2 \frac{\pi}{2(2n+1)} \end{array} \right.$$

$$y = \frac{\alpha}{\beta} = \frac{(2-\beta)^2}{\beta}$$

$$\text{si } n \text{ grand, } \beta \approx 4 \cdot \frac{\pi^2}{4(2n+1)^2} \approx \frac{\pi^2}{4n^2}$$

$$\Rightarrow \boxed{y \approx 1,6 n^2}$$

Système résolu

$$[A_1] [X] = [B]$$

$$\text{avec } B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

Le vecteur solution $[X]$ est égal à la première colonne de la matrice inverse, c'est-à-dire

$$[X] = \begin{bmatrix} n \\ n-1 \\ n-2 \\ \vdots \\ \vdots \\ 2 \\ 1 \end{bmatrix}$$

si n est la dimension du système.

Résultats obtenus

Les résultats sont indiqués dans le tableau ci-dessous. On peut y lire

- en première colonne → la dimension du système
- en deuxième colonne → $y = \frac{\alpha}{\beta}$ (le "nombre condition")
- en troisième colonne → le degré du polynôme $P_p(A)$ décomposé sous la forme $Q \times S = P$ (voir chapitre (1.2.2.2.) la signification de ces entiers)
- en quatrième colonne → $h_p = \frac{1}{T_p \left(\frac{\alpha + \beta}{\alpha - \beta} \right)}$
- en cinquième colonne → K nombre d'itérations principales effectuées
- en sixième colonne → $T = K \times P$ = le nombre total de produits du type $[A] [X]$

- en septième colonne $\rightarrow \max_j |x_k^j - x^j|$ avec
- $$\left\{ \begin{array}{l} x^j = \text{jème composante du vecteur exact } [X] \\ x_k^j = \text{jème composante du vecteur solution approché } [X_k] \end{array} \right.$$
- en huitième colonne $\rightarrow$ le nombre de chiffres significatifs exacts.

Remarque : Pour toutes ces résolutions, le vecteur de départ était le vecteur nul.
La méthode de Richardson est utilisée selon la technique R.A.C.

N	$y = \frac{\alpha}{\beta}$	$Q \times S = P$	h_p	K	$T = K \times P$	$\max_j x_j - x_j $	Nombre de chiffres exacts
20	678	$2 \times 24 = 48$	0,05	7	336	$0,2384 \cdot 10^{-6}$	7
		$5 \times 24 = 120$	0,000197	2	240	$0,7153 \cdot 10^{-6}$	7
		$6 \times 40 = 240$	$0,194 \cdot 10^{-7}$	1	240	$0,477 \cdot 10^{-6}$	7
50	4131	$5 \times 24 = 120$	0,0477	5	600	$0,267 \cdot 10^{-4}$	6
		$5 \times 40 = 200$	0,00396	3	600	$0,334 \cdot 10^{-5}$	7
		$6 \times 40 = 240$	0,00114	3	720	$0,953 \cdot 10^{-6}$	7
100	16370	$5 \times 24 = 120$	0,299	10	1200	$0,430 \cdot 10^{-3}$	5
		$6 \times 40 = 240$	0,0469	5	1200	$0,315 \cdot 10^{-4}$	6
200	65171	$6 \times 40 = 240$	0,298	9	2160	$0,348 \cdot 10^{-2}$	4

$$\text{Rappel : si } \forall i = 2, \dots, K \quad \frac{\|X_i - X_{i-1}\|}{\|X_{i-1} - X_{i-2}\|} \leq h_{\text{théorique}}$$

(ou du moins reste comparable à $h_{\text{théorique}}$).

La solution approchée $[X_K]$, est vis à vis de la solution exacte telle que

$$\left| \begin{array}{l} \frac{\|X - X_k\|}{\|X\|} \leq h^k \quad (\text{si } [X_0] = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}) \\ \|X - X_k\| \leq \|X_k - X_{k-1}\| \frac{h}{1-h} \end{array} \right.$$

Ces formules se sont révélées valables quelles que soient les dimensions des systèmes résolus.

Exemple

Résolution du système $N = 200$

$$K = 9$$

Pour $i = 2, 3, \dots, 9$, les quantités $\frac{\|X_i - X_{i-1}\|}{\|X_{i-1} - X_{i-2}\|}$ valaient respectivement 0,330 ; 0,300 ; 0,306 ; 0,300 ; 0,303 ; 0,300 ; 0,302 ; 0,301 ; et restaient donc comparables à $h_{\text{théorique}} = 0,298$.

$$\|X - X_9\| = 0,0359$$

$$\|X\| \simeq 1639,1$$

$$\Rightarrow \frac{\|X - X_9\|}{\|X\|} = 0,217 \cdot 10^{-4}$$

$$h = 0,298 \Rightarrow h^9 \simeq 0,185 \cdot 10^{-4}$$

La comparaison de ces deux quantités montre, que h^9 , s'il ne fournit pas exactement une majoration de l'erreur "relative", en donne au moins une approximation valable.

$$\|X_9 - X_8\| = 0,28148$$

$$\Rightarrow \|X_9 - X_8\| \frac{h}{1-h} \simeq 0,119$$

La deuxième formule d'erreur, donnant l'erreur sur les normes au sens absolu est bien vérifiée puisque

$$\|X - X_9\| \simeq 0,0359$$

Comparaison avec la méthode de Stiefel Frank (voir 1.2.3.2.)

Le système linéaire à 200 inconnues a aussi été résolu par la méthode de Stiefel Frank. La résolution a été plus rapide car l'erreur $\|X - X^*\| = 0,0359$ était atteinte en 1440 itérations de Stiefel Frank (au lieu de 2160 par la méthode de Richardson).

1. 4. 2. Matrice pleine n° 1

La matrice du deuxième système étudié est une matrice symétrique, dont les éléments, tous positifs, sont des entiers. Cette matrice appelée (réf. 11) "matrice inverse de Rutherford" est définie par

$$a_{ij} = 2 \text{ Min } (i, j) - 1$$

$$[A_2] = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 3 & 3 & \dots & 3 \\ 1 & 3 & 5 & \dots & 5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 3 & 5 & \dots & 2n-1 \end{bmatrix}$$

Inverse de $[A_2]$

$$[A_2]^{-1} = \frac{1}{2} \begin{bmatrix} 3 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \\ 0 & -1 & 2 & -1 & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}$$

Valeurs propres

Les valeurs propres de $[A_2]$ sont toutes positives, et, pour n très grand

$$y = \frac{\alpha}{\beta} \approx 1,6n^2$$

Système résolu

$$[A_2][X] = [B]$$

$$\text{avec } [B] = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Le vecteur solution de ce système est la première colonne de la matrice inverse de $[A_2]$

$$\text{et } [X] = \begin{bmatrix} 1,5 \\ -0,5 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

Résultats obtenus

Les systèmes ont été résolus pour des dimensions égales respectivement à 10, 20, 50 et 100 et les indications concernant ces résolutions sont groupées dans le tableau ci-dessous, les colonnes gardant les mêmes significations.

N	$y = \frac{\alpha}{\beta}$	$Q \times S = P$	h_p	K	$T = K \times P$	$\text{Max } x_{k_j} - x_j $	Nombre de chiffres exacts
10	~ 163	$4 \times 28 = 112$	$0,464. 10^{-7}$	1	112	$0,2386. 10^{-7}$	7
20	~ 624	$6 \times 40 = 240$	$0,892. 10^{-8}$	1	240	$0,486. 10^{-7}$	7
50	~ 4030	$6 \times 40 = 240$	$0,104. 10^{-2}$	2	480	$0,8792. 10^{-6}$	6
100	~ 16117	$6 \times 40 = 240$	0,04558	4	960	$0,274. 10^{-5}$	5

Pour toutes ces résolutions le vecteur initial était le vecteur nul.

Vérification des formules de contrôle d'erreur

Toutes ces résolutions se sont effectuées de façon conforme à la théorie, et les formules de majoration d'erreur se trouvaient être vérifiées ; à titre d'exemple : *Vérification de la majoration "relative" en norme*

- N = 50 K = 2 itérations principales de degré 240 ($[X_0] = [0]$)

$$\left. \begin{aligned} \frac{\|X_2 - X_1\|}{\|X_1 - X_0\|} &= 0,796. 10^{-3} \\ h_{\text{théor}} &= 0,1039. 10^{-2} \end{aligned} \right\} \Rightarrow h_{\text{eff}} < h_{\text{théo.}}$$

$$\Rightarrow \begin{array}{l} \boxed{h^2 = 0,108 \cdot 10^{-5}} \\ \|X - X_2\| = 0,119 \cdot 10^{-5} \\ \|X\| \quad 1,58 \end{array}$$

$$\Rightarrow \boxed{\frac{\|X - X_2\|}{\|X\|} = 0,753 \cdot 10^{-6}}$$

La relation $\frac{\|X - X_2\|}{\|X\|} \leq h^2$ est donc bien vérifiée.

- N = 100 K = 4 itérations principales de degré 240 ($[X_0] = [0]$)

$$\frac{\|X_2 - X_1\|}{\|X_1 - X_0\|} = 0,0336$$

$$\frac{\|X_3 - X_2\|}{\|X_2 - X_1\|} = 0,0409$$

$$\frac{\|X_4 - X_3\|}{\|X_3 - X_2\|} = 0,0431$$

La convergence s'est déroulée conformément aux prévisions théoriques puisque ces trois quantités sont toujours inférieures à h_{240} avec $h_{240} = 0,0456$

Nous vérifions encore que la majoration d'erreur relative sur les normes est justifiée.

car $h^4 \simeq 0,431 \cdot 10^{-5}$

et $\frac{\|X - X_4\|}{\|X\|} = 0,264 \cdot 10^{-5}$

1. 4. 3. Matrice pleine n° 2

La matrice que nous définissons ci-dessous a été introduite comme matrice test, car elle permet, en jouant sur un paramètre (R) de rendre la quantité α/β aussi grande que l'on veut, pour de petites dimensions de la matrice.

Cette matrice intervient dans des problèmes d'interpolation trigonométrique sur un demi-cercle de rayon R (voir II. 2.1).

Définition de $[A_4]$

$$[A_4]=\begin{bmatrix} 1 & R\cos \theta_1 & R^2 \cos 2\theta_1 & \dots & R^{n-1} \cos (n-1)\theta_1 \\ 1 & R\cos \theta_2 & R^2 \cos 2\theta_2 & \dots & R^{n-1} \cos (n-1)\theta_2 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & R\cos \theta_i & R^2 \cos 2\theta_i & \dots & R^{n-1} \cos (n-1)\theta_i \\ \dots & \dots & \dots & \dots & \dots \\ 1 & R\cos \theta_n & R^2 \cos 2\theta_n & \dots & R^{n-1} \cos (n-1)\theta_n \end{bmatrix}$$

$$\text{avec } \theta_i = \frac{(i-1) \pi}{n-1} \implies \begin{cases} \theta_1 = 0 \\ \theta_n = \pi \end{cases}$$

donc le terme général de $[A_4]$ est $a_{ij} = R^{j-1} \cos \left(\frac{(i-1) (j-1) \pi}{n-1} \right)$

$[A_4] = [B_4][C_4]$

avec $[B_4] =$

$$\begin{bmatrix} 1 & \cos \theta_1 & \dots & \cos (n-1)\theta_1 \\ 1 & \cos \theta_2 & \dots & \cos (n-1)\theta_2 \\ \dots & \dots & \dots & \dots \\ 1 & \cos \theta_i & \dots & \cos (n-1)\theta_i \\ \dots & \dots & \dots & \dots \\ 1 & \cos \theta_n & \dots & \cos (n-1)\theta_n \end{bmatrix}$$

$[C_4] =$

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & R & & \\ \vdots & & \ddots & \\ 0 & & & R^{n-1} \\ 0 & \dots & 0 & 0 \end{bmatrix}$$

Inverse de $[A_4]$

$$[A_4]^{-1} = [C_4]^{-1} [B_4]^{-1}$$

avec $[C_4]^{-1} =$

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1/R & & \\ & & 1/R^2 & \\ & & & \ddots \\ 0 & \dots & 0 & 1/R^{n-1} \end{bmatrix}$$

$$[B_4]^{-1} = \frac{1}{n-1} \begin{bmatrix} 1/2 & 1 & \dots & 1 & 1/2 \\ 1 & 2\cos\theta_2 & \dots & 2\cos\theta_{n-1} & \cos\theta_n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 2\cos i\theta_2 & \dots & 2\cos i\theta_{n-1} & \cos i\theta_n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 2\cos(n-2)\theta_2 & \dots & 2\cos(n-2)\theta_{n-1} & \cos(n-2)\theta_n \\ 1/2 & \cos(n-1)\theta_2 & \dots & \cos(n-1)\theta_{n-1} & \frac{\cos(n-1)\theta_n}{2} \end{bmatrix}$$

Système résolu :

Les valeurs propres de $[A_4]$ n'étant pas de même signe, le système résolu a été

$$[A_4^t] [A_4] [X] = [B]$$

le vecteur $[B]$, étant pour chaque essai calculé

de façon à rendre

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{bmatrix}$$

Résultats obtenus

Sont groupés dans ce tableau les indications concernant les résolutions faites pour $n=10$ et $n=20$ et pour respectivement $R=1,5$ et $R=1,25$. La technique de résolution est la technique R.A.C.

Pour chaque résolution le vecteur initial était le vecteur nul.

N	R	$y - \frac{\alpha}{\beta}$	degré	h	K	Nombre tot. de produits [A] [X]	$\text{Max}_j x_k^j - x^j $	Nombre mini. de chiffres exacts
10	1,5	1405	$5 \times 32 = 160$	0,00039	4	640	$0,36 \cdot 10^{-6}$	7
			$4 \times 40 = 160$	0,00039	4	640	$0,30 \cdot 10^{-6}$	7
			$5 \times 40 = 200$	0,000046	2	400	$0,53 \cdot 10^{-4}$	5
20	1,25	4400	$10 \times 24 = 240$	0,00144	2	480	0,00201	3

Comparaison avec la méthode de Stiefel-Frank

Ces deux systèmes ont été résolus par la méthode de Stiefel-Frank.

α) N = 10 R = 1,5

400 itérations de Stiefel-Frank ont été effectuées.

Nous comparons les résolutions effectuées selon cette méthode de Stiefel-Frank, et celles obtenues par la technique R.A.C. en itérant avec deux itérations principales de degré 200.

Si $[Y^j]$ est le jième vecteur itéré selon la méthode de Stiefel-Frank nous devrions avoir $[Y^{200}] = [X_1]$

- à la 200^{ième} itération de Stiefel-Frank -

la comparaison de
 $[Y^{200}]$ avec $[X]$ donne

$$\begin{aligned} \|Y^{200} - X\| &\simeq 0,00640 \\ \text{Max}_j |y_j^{200} - x_j| &\simeq 0,00527 \end{aligned}$$

et si $[X_1]$ est le 1er itéré (technique RAC)

$$\implies [X_1] \simeq [Y^{200}]$$

$$\begin{aligned} \|X_1 - X\| &\simeq 0,00647 \\ \text{Max}_j |x_{1j} - x_j| &\simeq 0,00537 \end{aligned}$$

- à la 400^{ième} itération de Stiefel-Frank -

$[Y^{400}]$ est vis à vis de $[X]$ tel que

$$\begin{aligned} \|Y^{400} - X\| &\simeq 0,0000526 \\ \text{et} \quad \text{Max}_j |y_j^{400} - x_j| &\simeq 0,0000517 \end{aligned}$$

en deux itérations, selon la technique RAC, c'est-à-dire en un temps équivalent (400 produits [A] [X]), le vecteur itéré $[X_2]$ est vis à vis de $[X]$

tel que

$$\begin{aligned} \|X_2 - X\| &= 0,0000541 \\ \text{Max}_j |x_{2j} - x_j| &= 0,0000532 \end{aligned}$$

Ces résultats prouvent que la convergence de la méthode de Stiefel-Frank est freinée par les erreurs de troncature. En effet si celles-ci n'influaient pas sur le déroulement de la convergence, $[Y^{400}]$, devrait être vis à vis de $[Y^{200}]$ tel

$$\text{que } \|Y_{400} - X\| \simeq \|Y_{200} - X\| \frac{h_{200}}{2 \cdot h_{200}^2}$$

$$\text{car } h_{400} \simeq \frac{h_{200}^2}{2 \cdot h_{200}^2} \simeq \frac{h_{200}^2}{2}$$

$$\text{or } \|Y_{400} - X\| = \|Y_{200} - X\| * 0,8 \cdot 10^{-2}$$

$$\text{et } h_{200} = 0,000046.$$

$$b) \underline{N = 20 \quad R = 1,25}$$

480 itérations de Stiefel-Frank ont été faites

- à la 240^{ième} itération -

$$\Rightarrow \left| \begin{array}{l} \|Y^{240} - X\| = 0,09221 \\ \text{Max}_j |y_j^{240} - x_j| = 0,07267 \end{array} \right.$$

- à la 480^{ième} itération -

$$\left| \begin{array}{l} \|Y^{480} - X\| = 0,004417 \\ \text{Max}_j |y_j^{480} - x_j| = 0,004406 \end{array} \right.$$

Résolution selon la technique R.A.C.

Les vecteurs $[X_1]$, et $[X_2]$ obtenus en des temps équivalents sont tels que

$$\left\{ \begin{array}{l} \|X_1 - X\| = 0,06545 \\ \text{Max}_j |x_{1j} - x_j| = 0,03347 \end{array} \right. \quad \text{avec} \quad \frac{\|X_2 - X_1\|}{\|X_1 - X_0\|} = 0,00120$$

$$\left\{ \begin{array}{l} \|X_2 - X\| = 0,002012 \\ \text{Max}_j |x_{2j} - x_j| = 0,002011 \end{array} \right. \quad h_{240} = 0,00144$$

Dans cette résolution, le calcul d'un polynôme $[P_{240}(A)]$ s'effectue donc mieux en utilisant ses racines (R.A.C.), qu'en utilisant la relation de récurrence.

- La technique R.E.C. a aussi été testée, avec deux polynômes de degré 240 (calculés donc par récurrence).

Les itérés obtenus que nous pouvons noter $[Z_1]$ et $[Z_2]$ devraient être théoriquement identiques respectivement à $[X_1]$ et $[X_2]$.

on obtient encore

$$\|Z_1 - X\| = 0,09221 \quad (\text{évident car } Z_1 = Y_{200})$$

$$\max_j |z_{1j} - x_j| = 0,07267$$

$$\|Z_2 - X\| = 0,004418$$

$$\max_j |z_{2j} - x_j| = 0,004408$$

Pour cet exemple, et contrairement aux prévisions, la technique R.A.C. s'est révélée être la meilleure. Ceci est néanmoins exceptionnel.

1. 4. 4. Conclusion

Pour les deux premiers systèmes résolus les erreurs d'arrondi paraissent négligeables. Cela est vraisemblablement dû au fait que pour le premier la matrice des coefficients est très "creuse", et pour le second son second membre.

La méthode de Stiefel-Frank, semble pour ce type de résolution plus efficace que la technique R.A.C. Mais si le gain en nombre d'itérations simples est dans le rapport 2/3, pour le premier exemple, le gain en temps machine est en fait moins sensible. Une itération de Stiefel-Frank, exige en effet un plus grand nombre d'opérations arithmétiques qu'une sous itération de la méthode R.A.C. (Pour s'en convaincre il suffit de comparer (XVIII - 2) page 33 et (X - 2) page 22).

Cette plus grande complexité dans le calcul des sous-itérés de la méthode de Stiefel-Frank, explique sans doute le phénomène qui intervenait pour la résolution du troisième système test, où le calcul d'un itéré principal au suivant, se faisait mieux en utilisant les racines (R.A.C.), que la relation de récurrence des polynômes $p_j(x)$.

Enfin nous signalerons que nous n'avons pas introduit pour ces systèmes une itération de la méthode semi-itérative de Tchebycheff, à la deuxième itération principale (R.A.C.).

1. 5. Compléments : Détermination des valeurs propres extrêmes - **Utilisation des polynômes de Tchebycheff pour la recherche** **de la valeur propre la plus petite en module.**

1. 5. 1. Nécessité d'une bonne connaissance des valeurs propres extrêmes

Lors de l'étude du polynôme $p_p(x)$, assurant pour un degré p la convergence la plus rapide de la suite vectorielle $[X_k] = [P_p(A)] [X_{k-1}] - [Q_{p-1}(A)] [B]$ vers $[A^{-1}] [B]$, nous avons souligné (1.1.5.5.) combien il importait de déterminer des quantités α et β qui soient aussi proches que possible de $|\lambda_1| = \max_i |\lambda_i|$ et de $\lambda_n = \min_i |\lambda_i|$

Les valeurs propres de $[A]$ n'étant strictement connues que dans des cas très exceptionnels, il faut pour déterminer α et β recourir à des procédés numériques.

Le plus simple et le plus rapide d'entre eux consisterait à majorer et minorer les valeurs propres par utilisation du théorème de Hadamard (réf. 9). L'expérience a montré que ce procédé n'était que très rarement applicable, les majorations et minorations qu'il fournit étant trop grossières.

Des techniques itératives sont alors conseillées. La méthode itérative classique (appelée parfois "méthode des puissances"), et que nous rappelons, a toujours permis l'obtention de la plus grande valeur propre en module, c'est-à-dire la quantité α , sans causer de difficultés majeures.

Son utilisation pour la recherche de β , s'est parfois révélée très coûteuse. Nous avons mis en oeuvre, à cet effet, une méthode itérative, utilisant les propriétés des polynômes de Tchebycheff qui a toujours permis d'approcher β avec une précision très suffisante.

1. 5. 2. Détermination numérique de α , valeur propre la plus grande en module, par méthode itérative.

Le procédé itératif, que nous rappelons, permet d'obtenir, avec son signe, la valeur propre la plus grande en module d'une matrice $[A]$, lorsque toutes ses valeurs propres sont réelles.

$[A]$ étant définie positive, ou définie négative, donc diagonalisable, ses vecteurs propres $[X_i]$ sont tous linéairement indépendants et forment une base.

Tout vecteur $[V_0]$, rapporté à cette base peut alors s'écrire :

$$[V_0] = c_1 [X_1] + c_2 [X_2] + \dots + c_n [X_n]$$

considérons la suite vectorielle de terme général

$$[V_p] = [A] [V_{p-1}]$$

$$\begin{aligned}
 [V_1] &= [A] [V_0] = c_1 [A] [X_1] + \dots + c_n [A] [X_n] = c_1 \lambda_1 [X_1] + \dots + c_n \lambda_n [X_n] \\
 [V_2] &= [A] [V_1] = c_1 \lambda_1 [A] [X_1] + \dots + c_n \lambda_n [A] [X_n] = c_1 \lambda_1^2 [X_1] + \dots + c_n \lambda_n^2 [X_n] \\
 (I-5) \quad \dots\dots\dots \\
 [V_p] &= [A] [V_{p-1}] = c_1 \lambda_1^{p-1} [A] [X_1] + \dots + c_n \lambda_n^{p-1} [A] [X_n] = c_1 \lambda_1^p [X_1] + \dots + c_n \lambda_n^p [X_n] \\
 [V_{p+1}] &= [A] [V_p] \dots\dots\dots = c_1 \lambda_1^{p+1} [X_1] + \dots + c_n \lambda_n^{p+1} [X_n]
 \end{aligned}$$

$[V_p]$ peut s'écrire :

$$[V_p] = \lambda_1^p (c_1 [X_1] + \left(\frac{\lambda_2}{\lambda_1}\right)^p c_2 [X_2] + \dots + \left(\frac{\lambda_n}{\lambda_1}\right)^p c_n [X_n]).$$

si les λ_i sont tels que

$$\begin{aligned}
 &|\lambda_1| > |\lambda_2| \geq \dots > |\lambda_n| \\
 \Rightarrow &\forall i > 1, \frac{|\lambda_i|}{|\lambda_1|} < 1
 \end{aligned}$$

$$\text{si } p \rightarrow \infty \quad \begin{cases} [V_p] \simeq \lambda_1^p c_1 [X_1] \\ [V_{p+1}] \simeq \lambda_1^{p+1} c_1 [X_1] \end{cases}$$

$$\Leftrightarrow \boxed{[V_{p+1}] \simeq \lambda_1 [V_p]} \quad (II-5)$$

Lorsque p augmente, le vecteur itéré de la suite (I - 5) tend à devenir parallèle au vecteur propre associé à la plus grande valeur propre en module de $[A]$, et deux itérés successifs tendent à être liés par la relation (II - 5), ce qui détermine λ_1 .

Pour rendre ce procédé utilisable sur ordinateur, une légère transformation de la suite (I - 5) est nécessaire, car si $|\lambda_1| > 1$, les λ_i^p intervenant dans $[V_p]$ risquent d'être rapidement très grands en module et d'atteindre ainsi la grandeur maximum permise par l'ordinateur utilisé.

Un facteur de normalisation est à cet effet introduit à chaque itération de (I - 5) afin de ramener toutes les composantes de l'itéré correspondant à des grandeurs correctes.

Le terme général de la suite devient alors

$$\boxed{[V_p] = [A] [V_{p-1}^*] \quad \text{où} \quad [V_{p-1}^*] = \frac{[V_{p-1}]}{r_{p-1}}}$$

, r_{p-1} , le facteur de normalisation valant par exemple $\|V_{p-1}\|$, où la plus grande composante de $[V_{p-1}]$

$$\text{alors si } p \rightarrow \infty \quad \left| \begin{array}{l} [V_p] \sim \frac{\lambda_1^p}{r_1 r_2 \dots r_{p-1}} \quad c_1[X_1] \\ [V_{p+1}] \sim \frac{\lambda_1^{p+1}}{r_1 r_2 \dots r_p} \quad c_1[X_1] \end{array} \right.$$

à la limite, deux itérés successifs sont liés par la relation

$$\boxed{[V_{p+1}] = \frac{\lambda_1}{r_p} [V_p]} \quad (\text{III} - 5)$$

$$\text{Il se déduit } r_{p+1} = \frac{\lambda_1}{r_p} \cdot r_p = \lambda_1$$

La suite scalaire des facteurs de normalisation r_i , converge donc vers λ_1 .

$$\boxed{\lim_{p \rightarrow \infty} r_p = \lambda_1}$$

d'où la détermination numérique de $\lambda_1 = \alpha$.

1. 5. 3. Détermination de β valeur propre la plus petite en module de $[A]$.

Remarque préliminaire : Nous plaçant dans le cas où les deux valeurs propres extrêmes sont de même signe, nous supposons ici par simple raison de commodité qu'elles sont toutes positives. Les méthodes restent néanmoins valables lorsque tous les λ_i sont négatifs.

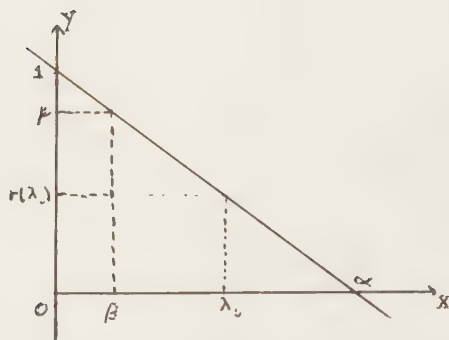
La méthode décrite en (1.5.2.) permet de calculer avec son signe la valeur propre la plus grande en module d'une matrice à valeurs propres toutes réelles.

Nous appliquons cette même technique pour rechercher la valeur propre la plus grande en module d'un polynôme matriciel de degré q , $[P_q(A)]$.

Nous savons que dans les conditions où nous nous plaçons, les valeurs propres de la matrice $[P_q(A)]$, sont les $p_q(\lambda_i)$, les vecteurs propres associés étant les $[X_i]$, vecteurs propres de $[A]$. Si le polynôme associé $p_q(x)$, est tel que $|p_q(\lambda_n)| > |p_q(\lambda_i)|$ pour $i = 1, \dots, n-1$ (avec $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \dots > |\lambda_n| = |\beta|$) les itérations (II - 5) ou (III - 5) appliquées à $[P_q(A)]$, permettront de calculer $\mu = p_q(\lambda_n) = p_q(\beta)$.

La valeur recherchée β sera alors la racine de l'équation $p_q(x) = \mu$.
Les méthodes décrites ci-dessous supposent α connu. ($|\alpha| \geq |\lambda_1|$).

1. 5. 3. 1. Détermination de β par méthode itérative classique.



Considérons le polynôme $r(x) = 1 - \frac{x}{\alpha}$, dont la représentation graphique est la droite joignant les points $(0, 1)$ et $(\alpha, 0)$. La fonction étant décroissante, il s'en suit que si $\lambda_1 > \lambda_2 > \lambda_3 \geq \dots > \lambda_n = \beta$ $r(\lambda_1) < r(\lambda_2) \leq \dots \leq r(\lambda_{n-1}) < r(\beta)$.

Deux vecteurs itérés successifs de la suite (II-5) appliquée à

la matrice $[B] = [R(A)] = [I] - \frac{[A]}{\alpha}$, seront donc liés à la limite par la relation

$$\lim_{p \rightarrow \infty} [V_{p+1}] = r(\beta) [V_p] \quad (\text{IV} - 5)$$

d'où
$$\underline{\underline{\beta = \alpha - r(\beta) \alpha}}$$

Remarque : Le choix du polynôme $r(x)$, ne nécessite pas à chaque itération la présence d'un facteur de normalisation, car sur $[\beta, \alpha]$ $r(x)$ est borné par 1.

Dans la majorité des problèmes cette méthode a permis d'obtenir une valeur approchant λ_n avec une précision suffisante. L'expérience a cependant montré que dans certains cas (en général lorsque β est très petit devant α , ce qui correspond aux problèmes difficiles à résoudre par la méthode de Richardson), la convergence était très longue. Parfois même, les itérations étaient telles que la valeur qui aurait dû normalement converger vers $r(\beta)$ prenait un comportement aléatoire autour d'une valeur moyenne, éloignée de $r(\beta)$.

Un exemple décrit à la fin de ce chapitre illustre ce phénomène.

Cet accident, dû aux erreurs de troncature, peut trouver son explication dans le fait, que le rapport $\frac{r(\lambda_{n-1})}{r(\lambda_n)}$ est très proche de 1. Or la rapidité avec laquelle $[V_p]$ devient parallèle au vecteur propre $[X_n]$ associé à $\beta = \lambda_n$, dépend de la rapidité avec laquelle tous les $\left(\frac{r(\lambda_i)}{r(\beta)}\right)^p$, $i = 1, \dots, n-1$, tendent vers zéro. Le processus est alors très long, et les erreurs de chute par leur cumul risquent de freiner et même d'altérer cette convergence. La méthode est alors à rejeter.

1. 5. 3. 2. Détermination de β dans les cas difficiles -

Méthode itérative avec utilisation des polynômes de Tchebycheff et "ajustage".

1°) Utilisation des polynômes de Tchebycheff: Justification

Le pième itéré de la suite $[V_i] = [R(A)] [V_{i-1}]$, avec $r(x) = 1 - \frac{x}{\alpha}$ est, si $[V_0] = c_1 [X_1] + c_2 [X_2] + \dots + c_n [X_n]$

$$[V_p] = r(\beta)^{\bar{p}} \left[c_1 \left(\frac{r(\lambda_1)}{r(\beta)} \right)^p [X_1] + \dots + c_i \left(\frac{r(\lambda_i)}{r(\beta)} \right)^p [X_i] + \dots + c_{n-1} \left(\frac{r(\lambda_{n-1})}{r(\beta)} \right)^p [X_{n-1}] + c_n [X_n] \right]$$

Supposons que nous appliquions les itérations (II - 5) à un polynôme matriciel de degré p , soit $[P_p(A)]$, avec $p_p(0) = 1$, et $|p_p(\beta)| > |p_p(\lambda_i)|$ $\forall i = 1, \dots, n-1$.

si $[W_1] = [P_p(A)] [V_0]$

$$[W_1] = p_p(\beta) \left[c_1 \frac{p_p(\lambda_1)}{p_p(\beta)} [X_1] + \dots + c_i \frac{p_p(\lambda_i)}{p_p(\beta)} [X_i] + \dots + c_{n-1} \frac{p_p(\lambda_{n-1})}{p_p(\beta)} [X_{n-1}] + c_n [X_n] \right]$$

$[W_1]$ est obtenu en un temps aussi rapide que $[V_p]$, car $p_p(x)$ est le produit de p polynômes du premier degré $(1 - \frac{x}{x_i})$, les x_i étant ses racines.

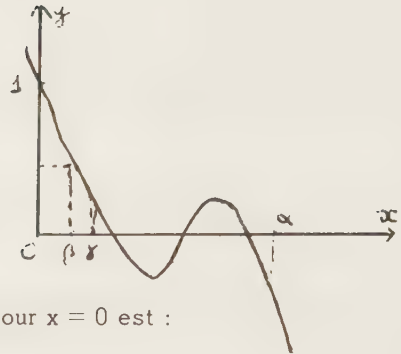
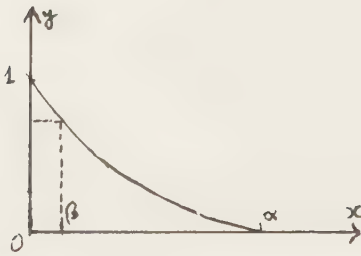
Plus le rapport $\frac{p_p(\lambda_{n-1})}{p_p(\beta)}$ sera inférieur à $(\frac{r(\lambda_{n-1})}{r(\beta)})^p$, plus proche sera

la direction de $[W_1]$ de celle du vecteur propre $[X_n]$, que ne l'était celle de $[V_p]$.

Nous trouvant devant un problème difficile, nous pouvons supposer que β est très proche de l'origine devant α .

Rendre le rapport $\frac{p_p(\lambda_{n-1})}{p_p(\beta)}$ minimum, revient alors à rechercher un polynôme $p_p(x)$, dont la dérivée, négative, soit la plus grande possible en valeur absolue à l'origine, donc au voisinage de β .

Nous savons qu'un polynôme de Tchebycheff $T_p(x)$, prend une valeur qui tend très rapidement vers l'infini (comme $1/2 \cdot e^{px}$), dès que $|x| \gg 1$. Un polynôme $p_p(x)$ peut donc être obtenu à partir d'un polynôme de Tchebycheff $T_p(x)$, en ramenant par un changement d'origine approprié l'intervalle $[-1, +1]$ à $[\gamma, \alpha]$, la quantité γ devant être supérieure à β .



Nous savons que ce polynôme, valant 1 pour $x = 0$ est :

$$p_p(x) = \frac{T_p\left(\frac{-2x + \alpha + \gamma}{\alpha - \gamma}\right)}{T_p\left(\frac{\alpha + \gamma}{\alpha - \gamma}\right)} = \frac{h_p}{2} [(u + \sqrt{u^2 - 1})^p + (u + \sqrt{u^2 - 1})^{-p}]$$

avec
$$\begin{cases} u = \frac{-2x + \alpha + \gamma}{\alpha - \gamma} \\ h_p = \frac{1}{T_p\left(\frac{\alpha + \gamma}{\alpha - \gamma}\right)} \end{cases}$$

Pour justifier ce choix, le tableau ci-dessous donne, pour diverses valeurs de α, β et p , les dérivées à l'origine prises par les deux polynômes

$$\left. \begin{aligned} & (r(x))^p \text{ et } p_p(x) \\ & \left| \begin{aligned} & \text{avec } \left(\frac{\partial (r(x))^p}{\partial x} \right)_{x=0} = - \frac{p}{\alpha} = \underline{d_1} \\ & \text{et } \left(\frac{\partial p_p(x)}{\partial x} \right) = - \frac{2p}{\alpha - \gamma} \operatorname{th} \left(p \operatorname{Argt} \operatorname{ch} \frac{\alpha + \gamma}{\alpha - \gamma} \right) \frac{1}{\sqrt{\left(\frac{\alpha + \gamma}{\alpha - \gamma} \right)^2 - 1}} = \underline{d_2} \end{aligned} \right. \end{aligned}$$

$\alpha = 5$

p	γ	t_1	d
1	0,5	0,2	0,1636
	1	0,2	0,3333
	2	-0,2	0,245
	3	-0,2	0,15
	4	-0,2	-0,2222
5	0,5	1	3,153
	1	-1	-2,236
	2	1	1,58114
	3	-1	1,231
	4	1	1,118
9	0,5	-1,8	5,692
	1	-1,8	4,0249
	2	-1,8	2,846
	3	1,8	2,324
	4	1,8	-2,0124

$$\alpha = 1000$$

p	γ	d_1	d_2
1	100	- 0,001	- 0,00182
	500	- 0,001	- 0,00133
	900	- 0,001	- 0,001052
5	100	- 0,005	- 0,0157
	500	- 0,005	- 0,00707
	900	- 0,005	- 0,00526
9	100	- 0,009	- 0,0285
	500	- 0,009	- 0,01273
	900	- 0,009	- 0,009486

Remarques : 1°) Pour un même degré, plus γ est proche de β , donc de l'origine, plus le rapport $\frac{d_2}{d_1}$ est élevé, donc plus le choix de ce polynôme est justifié.

2°) Si $\gamma \rightarrow \alpha \frac{d_2}{d_1} \rightarrow 1$, ce qui est logique car on vérifierait aisément que lorsque $\gamma \rightarrow \alpha \frac{d_2}{d_1} \rightarrow 1$, $p_p(x) \rightarrow (r(x))^p$.

Néanmoins $\forall \gamma \in [\beta, \alpha]$, la dérivée d_2 est plus élevée que d_1 , le rapport $\frac{d_2}{d_1}$ pouvant atteindre plusieurs unités.

3°) Plus le degré est élevé, plus le rapport $\frac{d_2}{d_1}$ est élevé.

Il y aura donc intérêt à itérer sur des polynômes de degrés assez élevés.

$$2^\circ) \text{ Etude de la suite vectorielle } [W_i] = \frac{[P_p(A)] [W_{i-1}]}{\|W_{i-1}\|} \quad (V - 5)$$

Nous retrouvons, une suite vectorielle identique à celle décrite en (1.5.2.1.), la matrice $[A]$ étant remplacée par $[P_p(A)]$, et le facteur de normalisation valant $\|W_{i-1}\| = r_{i-1}$.

Cette suite converge vers un vecteur de direction parallèle à celle du vecteur propre $[X_n]$ associé à la valeur propre $\beta = \lambda_n$ de $[A]$, et la suite scalaire $\|W_i\|$ a pour limite

$$\lim_{i \rightarrow \infty} \|W_i\| = p_p(\beta) = \|P_p(A)\|$$

La fonction $p_p(x)$ étant monotone décroissante entre 0 et γ , il s'en suit que l'équation $p_p(x) = \|P_p(A)\|$ admet une racine entre 0 et γ , qui est précisément β .

$$\text{Résolution de } p_p(x) = K = \|P_p(A)\|$$

Nous savons (1.1.5.3.) que si $c > 1$ l'équation $T_p(\omega) = c$ admet toujours pour racine $\omega = \frac{1}{2} [(c + \sqrt{c^2 - 1})^{1/p} + (c - \sqrt{c^2 - 1})^{1/p}]$ (si p est pair, présence d'une deuxième racine valant $-\omega$)

$$p_p(x) = K \iff T_p\left(\frac{-2x + \alpha + \gamma}{\alpha - \gamma}\right) = c \quad \text{avec } c = K/h_p > 1$$

$\forall_p$, la racine $x = \beta \in [0, \gamma]$ sera

$$\boxed{\beta = \frac{\alpha + \gamma}{2} - \frac{\alpha - \gamma}{4} [(c + \sqrt{c^2 - 1})^{1/p} + (c - \sqrt{c^2 - 1})^{1/p}]} \quad (\text{VI-5})$$

L'utilisation d'un tel polynôme $p_p(x)$ a toujours permis d'obtenir β avec une très bonne précision, et assez rapidement (voir exemple numérique en fin de chapitre).

3°) Recherche de β par méthode itérative utilisant les polynômes de Tchebycheff et avec "ajustage" de γ .

L'itération (V-5) définie par $[W_i] = \frac{[P_p(A)]}{\|W_{i-1}\|} [W_{i-1}]$ nécessite la détermination préalable d'une quantité γ , définie par $\beta < \gamma < \alpha$. Cette valeur, une fois choisie, est utilisée tout au long du processus itératif. Mais γ peut-être éloigné de β , et nous savons (voir ci-dessus 1°), que la pente à l'origine de $p_p(x)$ est d'autant plus forte en valeur absolue que γ est proche de β , par valeur supérieure.

Nous nous proposons ici, par un artifice reposant sur l'étude de la norme de l'itéré $[W_i]$, d'itérer, au stade suivant, à l'aide d'un polynôme construit, à partir d'une nouvelle valeur de γ , plus proche de β qu'elle ne l'était à l'itération précédente.

Les polynômes $p_p(x)$ varieront donc avec les itérations et, pour les personnaliser nous les repérerons par un indice (supérieur),

$$\text{ainsi } p_p^{(i)}(x) = \frac{T_p\left(\frac{-2x + \alpha + \gamma_i}{\alpha - \gamma_i}\right)}{T_p\left(\frac{\alpha + \gamma_i}{\alpha - \gamma_i}\right)} = h_p^{(i)} \cdot T_p\left(\frac{-2x + \alpha + \gamma_i}{\alpha - \gamma_i}\right)$$

$$\text{avec } h_p^{(i)} = \frac{1}{T_p\left(\frac{\alpha + \gamma_i}{\alpha - \gamma_i}\right)}$$

Détermination de γ_i en fonction de γ_{i-1}

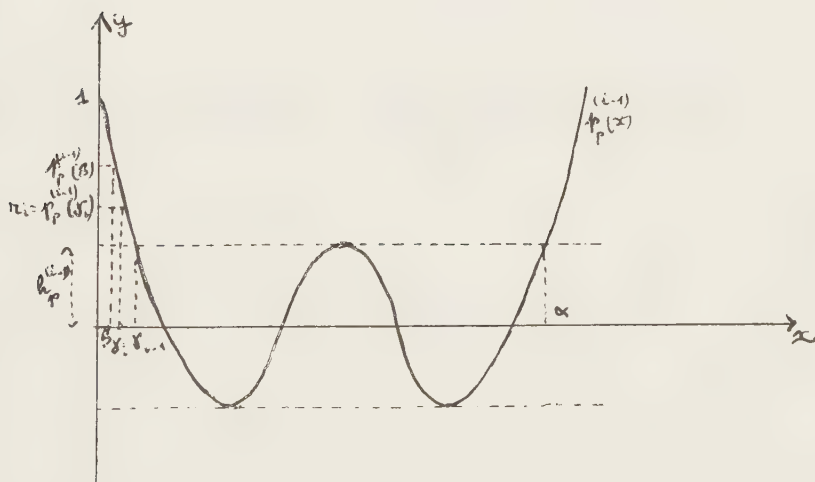
Supposons que le dernier vecteur calculé de la suite, soit

$$[W_i] = \frac{[P_p^{(i-1)}(A)] [W_{i-1}]}{r_{i-1}} \quad \text{avec} \quad r_{i-1} = \|W_{i-1}\|$$

nous savons que

$$\|W_i\| = \frac{\| [P_p^{(i-1)}(A)] [W_{i-1}] \|}{r_{i-1}} \leq \frac{\|P_p^{(i-1)}(A)\| \|W_{i-1}\|}{r_{i-1}}$$

$$\boxed{\|W_i\| \leq \|P_p^{(i-1)}(A)\| \quad \text{ou} \quad r_i \leq p_p^{(i-1)}(\beta)}$$



La comparaison de $r_i = \|W_i\|$ avec $h_p^{(i-1)} = p_p^{(i-1)}(\gamma_{i-1}) = |p_p^{(i-1)}(\alpha)|$ va nous permettre dans certains cas, de déterminer une valeur γ_i , plus proche de β que ne l'était γ_{i-1} .

a) $r_i > h_p^{(i-1)}$ (voir figure ci-dessus)

nous savons que $r_i \leq p_p^{(i-1)}(\beta) = \|P_p^{(i-1)}(A)\|$

$$\implies p_p^{(i-1)}(\beta) \geq r_i > h_p^{(i-1)} = p_p^{(i-1)}(\gamma_{i-1})$$

$p_p^{(i-1)}(x)$ étant une fonction monotone décroissante sur $[\beta, \gamma_{i-1}]$, la racine de l'équation $p_p^{(i-1)}(x) = r_i$, qui sera γ_i , est bien comprise entre β et γ_{i-1} , donc plus proche de β que ne l'était γ_{i-1} .

Cette racine définie par $\beta \leq \gamma_i < \gamma_{i-1}$, sera d'après (VI - 5)

$$\gamma_i = \frac{\alpha + \gamma_{i-1}}{2} - \frac{\alpha - \gamma_{i-1}}{4} [(C + \sqrt{C^2 - 1})^{1/p} + (C - \sqrt{C^2 - 1})^{1/p}] \quad (\text{VIII - 5})$$

avec $C = r_i / h_p^{(i-1)}$

b) si par contre $r_i \leq h_p^{(i-1)}$, nous ne pouvons pas de façon identique déterminer un nouveau $\gamma_i < \gamma_{i-1}$, L'itération suivante sera faite en utilisant la même valeur γ_{i-1} .

$$\gamma_i = \gamma_{i-1} \quad (\text{VIII - 5 bis})$$

en résumé si $[W_0] = [V_0] = c_1 [X_1] + \dots + c_n [X_n]$

et si γ_0 est une valeur initiale choisie proche de α ($\gamma_0 < \alpha$) et telle que $r_1 > h_p^{(0)}$ (sinon une autre valeur de γ_0 doit être choisie, encore plus proche de α).

La suite vectorielle $[W_{i+1}] = \frac{[P_p^{(i)}(A)] [W_i]}{r_i}$

(IX - 5)

avec

$$\left\{ \begin{array}{l} r_i = \|W_i\| \\ p_p^{(i)} = \frac{T_p\left(\frac{-2x + \alpha + \gamma_i}{\alpha - \gamma_i}\right)}{T_p\left(\frac{\alpha + \gamma_i}{\alpha - \gamma_i}\right)} \\ \gamma_i \text{ étant déterminé par } \left\{ \begin{array}{l} (\text{VIII-5}) \text{ si } r_i > h_p^{(i-1)} \\ (\text{VIII-5 bis}) \text{ si } r_i \leq h_p^{(i-1)} \end{array} \right. \end{array} \right.$$

convergera vers un vecteur parallèle au vecteur propre de $[A]$, associé à $\lambda_n = \beta$ soit $[X_n]$

$$\text{et } \lim_{k \rightarrow \infty} [W_{k+1}] = \frac{p^{(0)}(\beta) p^{(1)}(\beta) \dots p^{(k)}(\beta)}{r_0 r_1 r_2 \dots r_k} c_n [X_n]$$

avec à la limite
$$[W_{k+1}] = \frac{p_p^{(k)}(\beta)}{r_k} [W_k] \quad (X - 5)$$

1. 5. 4. Application numérique

Nous présentons ici un exemple de calcul des valeurs propres α et β par les méthodes décrites en (1.5.3.2.), c'est-à-dire en itérant sur des polynômes matriciels de Tchebycheff sans ou avec ajustage de la valeur γ à chaque itération.

Cet exemple a été choisi, car la méthode itérative classique (1.5.3.1.) était à rejeter.

Les méthodes décrites en 1.5.3.2. utilisent un certain nombre d'itérations de degré p , nécessitant chacune p produits du type $[A] [X]$. Le nombre total de tels produits permettra d'apprécier, en plus du gain en précision, le gain en rapidité de calculs.

1°) Matrice testée

$$[A] = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix}$$

Si n est la dimension de $[A]$, ses valeurs propres sont : $\lambda_i = 4 \sin^2 \frac{(2i-1)\pi}{4n+2}$

L'exemple décrit ci-dessous correspond à $n = 50$

$$\left\{ \begin{array}{l} \alpha = 3,9961312 \\ \beta = 0,00096744206. \end{array} \right.$$

2°) Résultats obtenus

a) par méthode itérative classique (1.5.3.1.)

au bout de	100 itérations	$\beta_{appr.}$	= 0,021507
	200 itérations	$\beta_{appr.}$	= 0,010072
	300 itérations	$\beta_{appr.}$	= 0,007331
	400 itérations	$\beta_{appr.}$	= 0,007412
	500 itérations	$\beta_{appr.}$	= 0,008767
	600 itérations	$\beta_{appr.}$	= 0,010357.

La méthode est à rejeter : Les erreurs de chute sont telles que la valeur obtenue, oscille dès la 200^{ième} itération et de façon aléatoire, autour d'une valeur moyenne 0,090.

Et quelque soit le nombre d'itération > 200 , $\longrightarrow 0,007 \leq \beta \leq 0,011$
 λ_n ne peut ainsi jamais être approché avec une précision suffisante.

b) Utilisation de polynômes de Tchebycheff sans ajustage

Une même valeur γ préalablement choisie est gardée tout au long des itérations (V - 5)

degré p	Valeur de γ utilisée	Nbre total d'itérations de degré p	Nbre total de produits [A] [X]	β
6	0,0060	82	492	0,000960
8	0,0089	55	440	0,000967423
10	0,015	53	530	0,00096744

c) Utilisation de polynômes de Tchebycheff avec ajustage

A chaque itération de la suite vectorielle (X - 5) est associée une valeur χ_i dépendant de l'itération précédente.

degré p	γ_0	Nbre total d'itérations de degré p	Nbre total de produits [A] [X]	β
4	$\infty/2$	75	300	0,00096754
6	$\infty/2$	39	234	0,00096745
8	$\infty/2$	23	184	0,00096753
10	$\infty/2$	16	160	0,00096746

DEUXIEME PARTIE

QUELQUES APPLICATIONS AUX EQUATIONS AUX DERIVEES PARTIELLES

Les méthodes itératives, en particulier celles décrites dans la première partie, ne semblent être présentées par les différents auteurs qui les recommandent (réf. [1], [2], [3]), que comme des méthodes de résolution des grands systèmes linéaires introduits lors de la résolution d'équations aux dérivées partielles par les différences finies (réf. [12]). Elles s'y prêtent effectivement bien et sont d'ailleurs les seules méthodes efficaces adaptées à ce genre de résolutions.

Contrairement aux méthodes dites directes, elles ne transforment pas la matrice du système, qui ne joue qu'un rôle d'opérande dans des produits du type $[Y] = [A] [X]$.

Dans le cas de résolutions aux différences finies cette matrice est très creuse c'est-à-dire ne comporte que très peu d'éléments non nuls (5 éléments au plus non nuls par ligne pour Δ , 13 pour $\Delta\Delta$). Elle n'est jamais explicitement construite et les produits $[A] [X]$ sont effectués à l'aide d'algorithmes qui ne mettent en jeu que ces éléments non nuls.

La matrice n'étant pas stockée en mémoire, des systèmes de dimensions très élevées peuvent ainsi être résolus (R.S. VARGA [3] cite un problème de 20 000 inconnues).

Dans cette partie nous présentons, comme première application, la résolution par les différences finies d'une équation aux dérivées partielles du type $\Delta\Delta \varphi = g(x, y)$ permettant de calculer des contraintes thermiques.

La méthode de Richardson, avec balayage, s'y est révélée très efficace. Un autre problème se rapportant au même opérateur $\Delta\Delta$ est en cours de résolution (réf. [13]) et sera publié ultérieurement.

Ces méthodes itératives peuvent aussi être intéressantes dans le cas de résolutions de systèmes pleins, de dimensions moyennes, (quelques centaines au plus). Les produits $[A] [X]$, nécessitent cette fois le stockage de la matrice. Si la capacité de l'ordinateur utilisé le permet, ce stockage pourra être fait en mémoire centrale. Dans le cas de petits ordinateurs, la matrice devra être rangée, lignes par lignes par exemple, sur mémoire périphérique (bandes, disques, tambour).

Le contrôle d'erreur pourra garantir la précision des résultats, qui dans les cas favorables seront excellents (6 à 7 chiffres significatifs exacts, pour chaque composante, ont souvent été obtenus pour des systèmes pleins de dimensions 50 environ).

Les méthodes directes sont le plus souvent utilisées pour ces types de systèmes. Si elles sont beaucoup plus rapides en temps d'exécution, aucune garantie n'est possible quand à la précision des résultats qu'elles fournissent. L'expérience a en effet montré que de telles méthodes n'étaient pas insensibles aux erreurs de chute, et conduisaient parfois à de mauvais résultats. Un mathématicien japonais (réf. [14]) démontre ainsi que la méthode de Gauss appliquée à la résolution d'un système linéaire très simple, car tridiagonal, peut donner selon la valeur des éléments non nuls de la matrice des résultats parfois excellents, mais aussi parfois totalement aberrants.

Comme application de la méthode de Richardson à la résolution des systèmes pleins, nous présentons dans le 2ème chapitre deux méthodes permettant de résoudre un problème de Dirichlet.

- par superposition finie de solutions harmoniques
- par utilisation de la théorie du potentiel.

Il est nécessaire pour la résolution d'un système linéaire par la méthode de Richardson, que les valeurs propres de la matrice soient toutes réelles et de même signe.

Nous avons été amenés, dans certaines des résolutions que nous présentons comme applications, à supposer à priori que les valeurs propres étaient effectivement réelles et de même signe. Ces hypothèses reposaient sur les propriétés des valeurs caractéristiques attachées aux opérateurs continus que l'on discrétisait. Le calcul approché des valeurs propres extrêmes α et β , par méthode itérative, puis la convergence de la méthode de Richardson ont toujours justifié ces hypothèses.

II-1. RESOLUTION D'UN PROBLEME D'ELASTICITE THERMIQUE

II - 1. 1. Problème de la lingotière - Mise en équation.

Des contraintes peuvent naître dans un corps, par suite de la non uniformité de la température. Ce phénomène se présente dans une lingotière pendant le remplissage.

Si $\mathcal{T}$ représente la dilatation ou variation unitaire de longueur au point (x, y, z) , les liaisons extérieures étant telles que cette variation ne provoque aucune réaction, les composantes du tenseur de déformation D' , due à la dilatation $\mathcal{T}$ sont :

$$\begin{array}{lcl} e'_1 & = & \mathcal{T} \quad g'_1 = 0 \\ e'_2 & = & \mathcal{T} \quad g'_2 = 0 \\ e'_3 & = & \mathcal{T} \quad g'_3 = 0 \end{array} \quad (I - 1)$$

Si ces conditions satisfont aux conditions d'intégrabilité, aucune contrainte ne prend naissance dans le corps, ce qui est le cas lorsque $\mathcal{T}$ est fonction linéaire de x, y, z .

Si les composantes de la déformation D' ne vérifient pas ces conditions d'intégrabilité, il se superposera à la déformation thermique D' une déformation élastique D'' , telle que la déformation résultante $D = D' + D''$ satisfasse aux conditions d'intégrabilité, c'est-à-dire telle que l'on puisse trouver en chaque point le déplacement (u, v, w) , et le tenseur contrainte $(n_1, n_2, n_3, t_1, t_2, t_3)$ vérifiant les équations d'équilibre.

II - 1. 1. 1. Enoncé du problème.

Soit un bloc de fonte parallélépipédique, de hauteur b , de longueur a et d'épaisseur infinie, contre lequel monte à une certaine vitesse de l'acier liquide à très haute température (1500°C). - (voir figure page suivante) -

Nous supposons

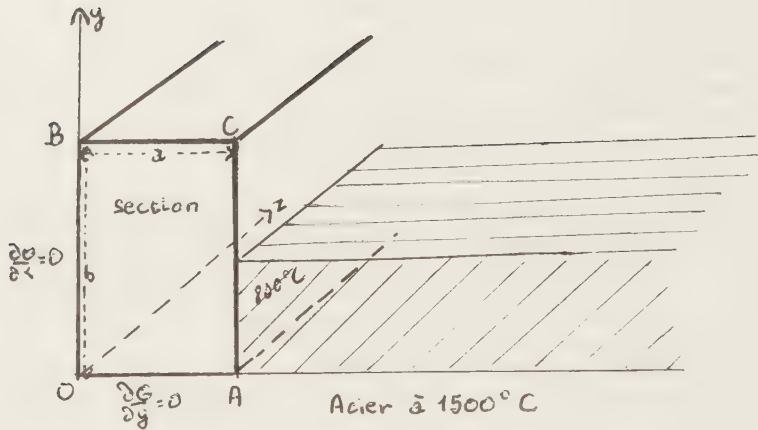
a) que la partie de la lingotière en contact avec l'acier en fusion reste à la température fixe de 800°C

b) qu'aucun flux de chaleur ne sort des faces OA_z et BO_z

$$\left(\frac{\partial \Theta}{\partial x} = 0 \text{ sur } OB, \quad \frac{\partial \Theta}{\partial y} = 0 \text{ sur } OA \right), \text{ ce qui revient}$$

à considérer $O\bar{A}$ et $O\bar{B}$, comme des axes de symétrie thermique.

c) qu'initialement, la lingotière était à 100°C .



- Figure 1 -

Des contraintes thermiques se créent au contact de l'acier liquide. La déformation qui en découle en un point (x, y, z) , ne dépend que de x et y , du fait de l'hypothèse d'infinité de l'épaisseur.

La déformation thermique qui s'introduit est donc plâne.

II - 1. 1. 2. Mise en équation.

Si Θ est la température en un point, la *dilatation* correspondante sera de la forme $\mathcal{T} = \alpha \Theta$ ($\alpha = \text{cste} = 10^{-5}$ par degré)

Les contraintes que nous proposons de déterminer devront satisfaire, comme pour tout problème d'élasticité [réf. n° 15]

- 1°) Aux équations d'équilibre
- 2°) Aux conditions aux limites
- 3°) A la condition d'intégrabilité

a) Les composantes de la déformation valent ici :

$$\begin{aligned}
 e_1 - \frac{\partial u}{\partial x} &= -\frac{1}{E} [n_1 - \sigma(n_2 + n_3)] + \mathcal{T} & g_1 &= \frac{1}{2} \left[\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \right] = 0 \\
 e_2 - \frac{\partial v}{\partial y} &= -\frac{1}{E} [n_2 - \sigma(n_1 + n_3)] + \mathcal{T} & g_2 &= \frac{1}{2} \left[\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \right] = 0 \\
 e_3 - \frac{\partial w}{\partial z} &= -\frac{1}{E} [n_3 - \sigma(n_1 + n_2)] + \mathcal{T} = 0 & g_3 &= \frac{1}{2} \left[\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right] = -\frac{(1+\sigma)}{E} t_3
 \end{aligned}
 \tag{II - 1}$$

u, v, w étant les déplacements suivant Ox, Oy, Oz . e_3, g_1 et g_2 sont nuls car u et v sont indépendants de z et w est nul.

$$e_3 = 0 \implies \boxed{n_3 = \sigma(n_1 + n_2) + E\mathcal{T}}$$

alors (II - 1) ~~(III - 1)~~ donnant le tenseur déformation en fonction de n_1, n_2 et t_3 .

$$\boxed{\begin{aligned} e_1 &= -\frac{(1+\sigma)}{E} [(1-\sigma)n_1 - \sigma n_2] + (1+\sigma)\mathcal{T} & g_1 &= 0 \\ e_2 &= -\frac{(1+\sigma)}{E} [(1-\sigma)n_2 - \sigma n_1] + (1+\sigma)\mathcal{T} & g_2 &= 0 \\ e_3 &= 0 & g_3 &= -\frac{(1+\sigma)}{E}t_3 \end{aligned}} \quad (\text{III} - 1)$$

NB : E est le module d'élasticité longitudinale, ou module d'Young ($F \cdot L^2$)
 σ est un coefficient sans dimension, appelé coefficient de Poisson.

b) Composantes du tenseur contrainte.

$$\begin{aligned} &\begin{cases} e_1 = -\frac{(1+\sigma)}{E} [(1-\sigma)n_1 - \sigma n_2] + (1+\sigma)\mathcal{T} \\ e_2 = -\frac{(1+\sigma)}{E} [(1-\sigma)n_2 - \sigma n_1] + (1+\sigma)\mathcal{T} \end{cases} \\ \implies &\begin{cases} (1-\sigma)n_1 - \sigma n_2 = -\frac{E}{(1+\sigma)}e_1 + E\mathcal{T} \\ -\sigma n_1 + (1-\sigma)n_2 = -\frac{E}{(1+\sigma)}e_2 + E\mathcal{T} \end{cases} \end{aligned}$$

l'élimination de n_1 , puis celle de n_2 , permettent de donner les valeurs de n_1, n_2 , puis $n_3 = \sigma(n_1 + n_2) + E\mathcal{T}$ en fonction de e_1 et e_2 : D'ou le tenseur contrainte :

$$\boxed{\begin{aligned} n_1 &= -\frac{E}{(1-2\sigma)} \left[\frac{\sigma e_2}{(1+\sigma)} + \frac{(1-\sigma)e_1}{(1+\sigma)} - \mathcal{T} \right] & t_1 &= 0 \\ n_2 &= -\frac{E}{(1-2\sigma)} \left[\frac{\sigma e_1}{(1+\sigma)} + \frac{(1-\sigma)e_2}{(1+\sigma)} - \mathcal{T} \right] & t_2 &= 0 \\ n_3 &= -\frac{E}{(1-2\sigma)} \left[\frac{(\sigma e_1 + e_2)}{(1+\sigma)} - \mathcal{T} \right] & t_3 &= -\frac{E}{(1+\sigma)}g_3 \end{aligned}} \quad (\text{IV} - 1)$$

Nous pouvons aussi faire intervenir la dilatation cubique Θ

$$\Theta = e_1 + e_2 + e_3 = e_1 + e_2$$

$$\begin{aligned} n_1 &= -\frac{E}{(1-2\sigma)} \left[\frac{\sigma \Theta}{(1+\sigma)} + \frac{(1-2\sigma)}{(1+\sigma)} e_1 - \mathcal{T} \right] & t_1 &= 0 \\ n_2 &= -\frac{E}{(1-2\sigma)} \left[\frac{\sigma \Theta}{(1+\sigma)} + \frac{(1-2\sigma)}{(1+\sigma)} e_2 - \mathcal{T} \right] & t_2 &= 0 \\ n_3 &= -\frac{E}{(1-2\sigma)} \left[\frac{\sigma \Theta}{(1+\sigma)} - \mathcal{T} \right] & t_3 &= -\frac{E}{(1+\sigma)} g_3 \end{aligned} \quad (V-1)$$

Mais nous ne considérons qu'un problème plan, d'où les caractéristiques qui restent utiles :

<i>Tenseur déformation</i>	<i>Tenseur contrainte</i>
$\begin{cases} e_1 = -\frac{(1+\sigma)}{E} [(1-\sigma) n_1 - \sigma n_2] + (1+\sigma) \mathcal{T} \\ e_2 = -\frac{(1+\sigma)}{E} [(1-\sigma) n_2 - \sigma n_1] + (1+\sigma) \mathcal{T} \\ g_3 = -\frac{(1+\sigma)}{E} [t_3] \end{cases}$	$\begin{cases} n_1 = -\frac{E}{(1-2\sigma)} \left[\frac{\sigma e_2}{(1+\sigma)} + \frac{(1-\sigma)e_1}{(1+\sigma)} - \mathcal{T} \right] \\ n_2 = -\frac{E}{(1-2\sigma)} \left[\frac{\sigma e_1}{(1+\sigma)} + \frac{(1-\sigma)e_2}{(1+\sigma)} - \mathcal{T} \right] \\ t_3 = -\frac{E}{(1+\sigma)} g_3 \end{cases} \quad (VI-1)$

Ces caractéristiques de la déformation doivent vérifier

- 1°) Les équations d'équilibres
- 2°) La condition d'intégrabilité
- 3°) Les conditions aux limites.

1°) *Equations d'équilibres.*

Elles se réduisent dans ce cas de déformation plane, les forces de volume étant supposées nulles, à :

$$\begin{cases} \frac{\partial n_1}{\partial x} + \frac{\partial t_3}{\partial y} = 0 \\ \frac{\partial t_3}{\partial x} + \frac{\partial n_2}{\partial y} = 0 \end{cases}$$

Si l'on peut trouver une fonction Ψ (fonction d'Airy), telle que :

$$\boxed{n_1 = \frac{\partial^2 \Psi}{\partial y^2} \quad n_2 = \frac{\partial^2 \Psi}{\partial x^2} \quad \text{et} \quad t_3 = -\frac{\partial^2 \Psi}{\partial x \partial y}} \quad (\text{VII} - 1)$$

ces conditions seront réalisées.

2°) La condition d'intégrabilité

est réduite à

$$\boxed{2 \frac{\partial^2 g_3}{\partial x \partial y} = \frac{\partial^2 e_1}{\partial y^2} + \frac{\partial^2 e_2}{\partial x^2}}$$

$$\frac{\partial^2 g_3}{\partial x \partial y} = -\frac{(1+\sigma)}{E} \frac{\partial^2 t_3}{\partial x \partial y} = -\frac{(1+\sigma)}{E} \frac{\partial^4 \Psi}{\partial x \partial y}$$

$$\frac{\partial^2 e_1}{\partial y^2} = -\frac{(1+\sigma)}{E} [(1-\sigma) \frac{\partial^2 n_1}{\partial y^2} - \sigma \frac{\partial^2 n_2}{\partial y^2}] + (1+\sigma) \frac{\partial^2 \mathcal{T}}{\partial y^2}$$

$$= -\frac{(1+\sigma)}{E} [(1-\sigma) \frac{\partial^4 \Psi}{\partial y^4} - \sigma \frac{\partial^4 \Psi}{\partial x^2 \partial y^2}] + (1+\sigma) \frac{\partial^2 \mathcal{T}}{\partial y^2}$$

$$\frac{\partial^2 e_2}{\partial x^2} = -\frac{(1+\sigma)}{E} [(1-\sigma) \frac{\partial^2 n_2}{\partial x^2} - \sigma \frac{\partial^2 n_1}{\partial x^2}] + (1+\sigma) \frac{\partial^2 \mathcal{T}}{\partial x^2}$$

$$= -\frac{(1+\sigma)}{E} [(1-\sigma) \frac{\partial^4 \Psi}{\partial x^4} - \sigma \frac{\partial^4 \Psi}{\partial x^2 \partial y^2}] + (1+\sigma) \frac{\partial^2 \mathcal{T}}{\partial x^2}$$

La condition d'intégrabilité devient :

$$2 \frac{\partial^4 \Psi}{\partial x^2 \partial y^2} + (1-\sigma) \frac{\partial^4 \Psi}{\partial y^4} - \sigma \frac{\partial^4 \Psi}{\partial x^2 \partial y^2} + (1-\sigma) \frac{\partial^4 \Psi}{\partial x^4} - \sigma \frac{\partial^4 \Psi}{\partial x^2 \partial y^2} - E \left[\frac{\partial^2 \mathcal{T}}{\partial y^2} + \frac{\partial^2 \mathcal{T}}{\partial x^2} \right] = 0$$

$$(1-\sigma) \left[\frac{\partial^4 \Psi}{\partial x^4} + \frac{\partial^4 \Psi}{\partial y^4} + 2 \frac{\partial^4 \Psi}{\partial x^2 \partial y^2} \right] = E \left[\frac{\partial^2 \mathcal{T}}{\partial y^2} + \frac{\partial^2 \mathcal{T}}{\partial x^2} \right]$$

$$\Rightarrow \Delta \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} \right) = \frac{E}{1-\sigma} \left(\frac{\partial^2 \mathcal{T}}{\partial x^2} + \frac{\partial^2 \mathcal{T}}{\partial y^2} \right)$$

$$\boxed{\Delta \Delta \Psi = \frac{E}{1-\sigma} \Delta \mathcal{T}} \quad (\text{VIII} - 1)$$

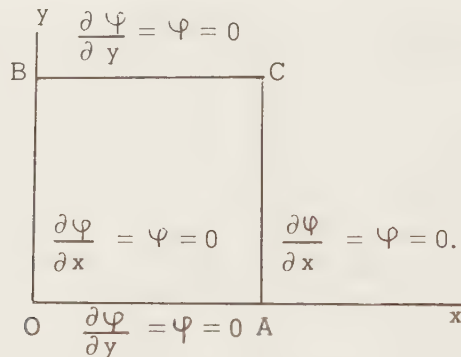
3°) Conditions aux limites.

Si α et β sont les composantes de la normale extérieure à la surface externe, elles s'écrivent :

$$\begin{cases} \alpha n_1 + \beta t_3 = -\bar{X} \\ \alpha t_3 + \beta n_2 = -\bar{Y} \end{cases}$$

Les forces de surface $\bar{X}$ et $\bar{Y}$ sont nulles

$$\Rightarrow \begin{cases} \alpha n_1 = -\beta t_3 \\ \alpha t_3 = -\beta n_2 \end{cases} \quad \text{ou} \quad \begin{cases} \alpha \frac{\partial^2 \varphi}{\partial y^2} = \beta \frac{\partial^2 \varphi}{\partial x \partial y} \\ \alpha \frac{\partial^2 \varphi}{\partial x \partial y} = \beta \frac{\partial^2 \varphi}{\partial x^2} \end{cases}$$



le contours étant rectangulaire, elles se réduisent à :

$$\begin{cases} \text{soit } \frac{\partial^2 \varphi}{\partial y^2} = 0 \text{ et } \frac{\partial^2 \varphi}{\partial x \partial y} = 0 & \text{sur OB et AC} \\ \text{soit } \frac{\partial^2 \varphi}{\partial x^2} = 0 \text{ et } \frac{\partial^2 \varphi}{\partial x \partial y} = 0 & \text{sur OA et BC} \end{cases}$$

par exemple

et

$\varphi = 0$ et $\frac{\partial \varphi}{\partial x} = 0$	pour $x = 0$ et $x = a$
$\varphi = 0$ et $\frac{\partial \varphi}{\partial y} = 0$	pour $y = 0$ et $y = b$

répondront au problème.

En conclusion :

Le problème posé sera résolu si nous réunissons à trouver une fonction φ telle que :

$$\Delta \Delta \varphi = \frac{E}{1 - \sigma} \Delta \mathcal{T}$$

et satisfaisant aux conditions aux limites

(IX - 1)

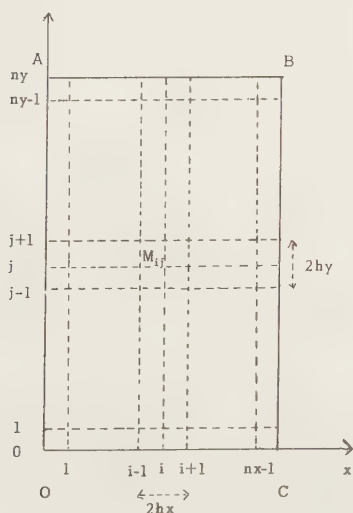
$$\varphi = 0 \quad \text{et} \quad \frac{\partial \varphi}{\partial x} = 0 \quad \forall y, \quad \text{pour } x = 0 \quad \text{et} \quad x = a$$

$$\varphi = 0 \quad \text{et} \quad \frac{\partial \varphi}{\partial y} = 0 \quad \forall x, \quad \text{pour } y = 0 \quad \text{et} \quad y = b$$

II - 1. 2. Problème de la lingotière - Résolution numérique.

La méthode utilisée pour résoudre numériquement ce problème a été la méthode classique et élémentaire des différences finies, méthode qui consiste à discrétiser les opérateurs qui interviennent dans l'équation. Elle conduit toujours à la construction d'un système linéaire dont le vecteur solution représente les valeurs de la fonction inconnue en un nombre élevé de points préalablement choisis.

II - 1. 2. 1. Construction d'un système linéaire.



- Figure 2 -

Soit OABC la section de la lingotière. Définissons un réseau à mailles rectangulaires, de mêmes cotés (h_x suivant Ox , h_y suivant Oy), et comportant au total (n_x+1) (n_y+1) noeuds M_{ij} , $\left\{ \begin{array}{l} i = 0, \dots, n_x \\ j = 0, \dots, n_y \end{array} \right.$

dont les coordonnées sont

$$\left\{ \begin{array}{l} x_i = i \cdot h_x \\ y_j = j \cdot h_y \end{array} \right.$$

Les inconnues du problème sont les valeurs de $\varphi(x, y)$ aux (n_x-1) (n_y-1) noeuds intérieurs du réseau, φ étant nul au pourtour par hypothèse.

Soit $\varphi_{ij} = \varphi(x_i, y_j)$

$$\text{avec} \quad \Delta \Delta \varphi_{ij} = \frac{E}{1 - \sigma} (\Delta \theta)_{ij}$$

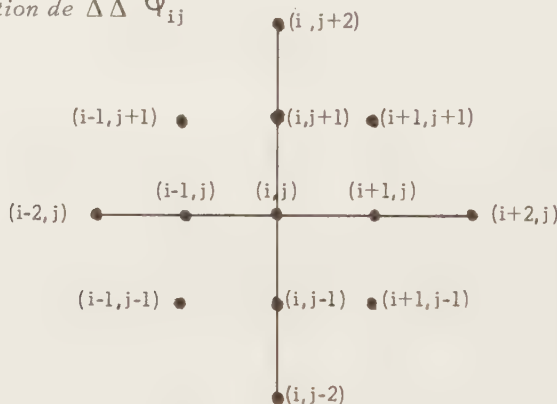
1°) Approche de $(\Delta \mathcal{T}_{ij})$

Les températures θ_{ij} données en un nombre très élevé de points (réseau à mailles carrées de 0,5 cm) ont permis le calcul une fois pour toutes de $(\Delta \mathcal{T})_{ij}$, en ces mêmes noeuds, en utilisant la discrétisation bien connue de l'opérateur Δ .

$$\Delta \mathcal{T}_{ij} = \frac{\mathcal{T}_{i+1,j} + \mathcal{T}_{i-1,j} - 2 \mathcal{T}_{ij}}{hx^2} + \frac{\mathcal{T}_{i,j+1} + \mathcal{T}_{i,j-1} - 2 \mathcal{T}_{ij}}{hy^2}$$

si $\mathcal{T}_{ij} = \propto \theta_{ij}$, θ_{ij} étant la température au noeud M_{ij} .

2°) Discrétisation de $\Delta \Delta \mathcal{Q}_{ij}$



$$\Delta \mathcal{Q}_{ij} = \frac{\mathcal{Q}_{i+1,j} + \mathcal{Q}_{i-1,j} - 2 \mathcal{Q}_{ij}}{hx^2} + \frac{\mathcal{Q}_{i,j+1} + \mathcal{Q}_{i,j-1} - 2 \mathcal{Q}_{ij}}{hy^2}$$

mais l'opérateur Δ est linéaire, $\implies \Delta \Delta \mathcal{Q}_{ij} = \Delta [\Delta \mathcal{Q}_{ij}]$

$$\implies \Delta \Delta \mathcal{Q}_{ij} = \frac{\Delta \mathcal{Q}_{i+1,j} + \Delta \mathcal{Q}_{i-1,j} - 2 \Delta \mathcal{Q}_{ij}}{hx^2} + \frac{\Delta \mathcal{Q}_{i,j+1} + \Delta \mathcal{Q}_{i,j-1} - 2 \Delta \mathcal{Q}_{ij}}{hy^2}$$

ce qui donne :

$$\Delta \Delta \mathcal{Q}_{ij} \simeq \varepsilon \mathcal{Q}_{ij} + c [\mathcal{Q}_{i,j+1} + \mathcal{Q}_{i,j-1}] + b [\mathcal{Q}_{i+1,j} + \mathcal{Q}_{i-1,j}] + \\ a [\mathcal{Q}_{i+2,j} + \mathcal{Q}_{i-2,j}] + d [\mathcal{Q}_{i,j+2} + \mathcal{Q}_{i,j-2}] + \\ e [\mathcal{Q}_{i+1,j+1} + \mathcal{Q}_{i+1,j-1} + \mathcal{Q}_{i-1,j+1} + \mathcal{Q}_{i-1,j-1}]$$

avec

$$\varepsilon = \frac{6}{hx^4} + \frac{6}{hy^4} + \frac{8}{hx^2hy^2}$$

$$c = -\frac{4}{hy^2} \left(\frac{1}{hx^2} + \frac{1}{hy^2} \right)$$

$$b = -\frac{4}{hx^2} \left(\frac{1}{hx^2} + \frac{1}{hy^2} \right)$$

$$a = \frac{1}{hx^4}$$

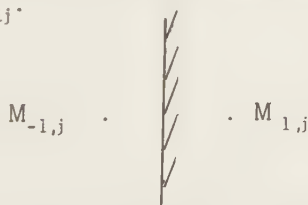
$$d = \frac{1}{hy^4}$$

$$e = \frac{2}{hx^2 hy^2}$$

La présence des termes en $\varphi_{i+2,j}$, $\varphi_{i-2,j}$, $\varphi_{i,j+2}$, et $\varphi_{i,j-2}$ implique que ce développement n'est pas valable pour les noeuds limitrophes (il ferait alors intervenir des noeuds extérieurs au domaine).

Pour ces points, nous tiendrons compte de la présence de dérivées normales nulles de $\varphi(x, y)$ à la frontière, ce qui modifiera alors légèrement l'approche de $\Delta \Delta \varphi_{ij}$.

Exemple :



$$\varphi(1,j) = \varphi(0,j) + hx \varphi'_x(0,j) + \frac{hx^2}{2!} \varphi''_{x^2}(0,j) + \varepsilon \left(\frac{hx^3}{3!} \right)$$

$$\varphi(-1,j) = \varphi(0,j) - hx \varphi'_x(0,j) + \frac{hx^2}{2!} \varphi''_{x^2}(0,j) + \varepsilon \left(\frac{hx^3}{3!} \right)$$

$\implies$

de même $\left\{ \begin{array}{l} \varphi(1,j) \simeq \varphi(-1,j) \\ \varphi(nx-1,j) \simeq \varphi(nx+1,j) \\ \varphi(i,1) \simeq \varphi(i,-1) \\ \varphi(i,ny-1) \simeq \varphi(i,ny+1) \end{array} \right\}$

Ainsi, par exemple, pour $i = 1$

- si $j \neq 1$ et $j \neq ny-1$

$$\Delta \Delta \varphi_{1j} = (\varepsilon + \alpha) \varphi_{1j} + c [\varphi_{1,j+1} + \varphi_{1,j-1}] + b \varphi_{2,j} + \alpha \varphi_{3,j} + d [\varphi_{1,j+\frac{1}{2}} \varphi_{1,j-2}] + e [\varphi_{2,j+1} + \varphi_{2,j-1}]$$

- si $j = 1$

$$\Delta \Delta \varphi_{1,1} = (\varepsilon + \alpha + d) \varphi_{1,1} + c \varphi_{1,2} + b \varphi_{2,1} + \alpha \varphi_{3,1} + d \varphi_{1,3} + e \varphi_{2,2}$$

- si $j = ny-1$

$$\Delta \Delta \varphi_{1,ny-1} = (\varepsilon + \alpha + d) \varphi_{1,ny-1} + c \varphi_{1,ny-2} + b \varphi_{2,ny-1} + \alpha \varphi_{3,ny-1} + d \varphi_{1,ny-3} + e \varphi_{2,ny-2}$$

Le système linéaire à résoudre est obtenu en écrivant qu'aux n noeuds intérieurs (si $n = (nx-1)(ny-1)$) la relation $\Delta \Delta \varphi_{ij} = \frac{E}{1-\sigma} \Delta \mathcal{T}_{ij}$ est vérifiée

$$\Rightarrow \boxed{[M] [\varphi] = [H]}$$

La matrice $[M]$ peut être partitionnée en $(nx-1) \times (nx-1)$ "sous matrices" carrées d'ordre $(ny-1)$.

Il en est de même pour les vecteurs colonnes $[\varphi]$ et $[H]$, qui peuvent être décomposés en $(nx-1)$ "sous vecteurs" à $(ny-1)$ composantes.

$$\begin{bmatrix} [A^*] [B] [C] [0] & . & . & . & [0] \\ [B] [A] [B] [C] [0] & . & . & . & . \\ [C] [B] [A] [B] [C] [0] & . & . & . & . \\ [0] [C] & . & . & . & [0] & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & [C] [B] [A] [B] & . & . \\ . & . & . & . & [0] [C] [B] [A^*] & . & . \end{bmatrix} \begin{bmatrix} [\psi_1] \\ [\psi_2] \\ . \\ . \\ . \\ . \\ [\psi_{nx-2}] \\ [\psi_{nx-1}] \end{bmatrix} = \begin{bmatrix} [H_1] \\ [H_2] \\ . \\ . \\ . \\ . \\ . \\ [H_{nx-1}] \end{bmatrix}$$

$$\text{avec } [A^*] = \begin{bmatrix} (\varepsilon + \alpha + d) & c & d & . & . & 0 \\ c & (\varepsilon + \alpha) & c & d & . & 0 \\ d & c & (\varepsilon + \alpha) & c & d & 0 \\ . & . & . & . & . & . \\ . & . & d & c & (\varepsilon + \alpha) & c \\ 0 & . & 0 & d & c & (\varepsilon + \alpha + d) \end{bmatrix}$$

$$[A] = \begin{bmatrix} (\epsilon + d) & c & d & 0 & 0 & 0 \\ c & \epsilon & c & d & 0 & . \\ d & c & \epsilon & c & d & . \\ . & . & . & . & . & d \\ . & . & . & . & \epsilon & c \\ 0 & . & . & d & c & (\epsilon + d) \end{bmatrix}$$

$$[B] = \begin{bmatrix} b & e & 0 & . & . & . & . & 0 \\ e & b & e & 0 & . & . & . & 0 \\ 0 & e & b & e & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & 0 & e & b & e \\ 0 & . & . & . & . & 0 & e & b \end{bmatrix}$$

$$[C] = \begin{bmatrix} \alpha & 0 & . & . & 0 \\ 0 & . & . & . & . \\ . & . & . & . & . \\ . & . & . & \alpha & 0 \\ 0 & . & . & 0 & \alpha \end{bmatrix} \quad \text{et } [0] = \begin{bmatrix} 0 & . & 0 \\ . & . & . \\ 0 & . & 0 \end{bmatrix}$$

$$\text{et } [\psi_i] = \begin{bmatrix} \psi_{i1} \\ \psi_{i2} \\ . \\ . \\ . \\ \psi_{i,ny-1} \end{bmatrix} \quad [H_i] = \frac{E_\infty}{1 - \sigma} \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \\ . \\ . \\ . \\ \theta_{i,ny-1} \end{bmatrix}$$

La matrice $[M]$, comporte au plus 13 éléments non nuls par ligne.

II - 1. 2. 2. Calcul des contraintes.

Nous supposons résolu le système $[M] [\varphi] = [H]$.

La valeur prise par la fonction $\varphi(x, y)$ est donc connue de façon discrète aux $n = (nx-1) \cdot (ny-1)$ noeuds intérieurs du réseau défini par $(hx$ et $hy)$ et

$$\varphi(x_i, y_j) = \varphi(i.hx, j.hy) = \varphi_{ij}$$

$$\left| \begin{array}{l} i = 1, \dots nx-1 \\ j = 1, \dots ny-1 \end{array} \right.$$

Aux noeuds du réseau appartenant à la frontière, $\varphi(x, y)$ est par définition nul.

$$\varphi(x_i, y_j) = \varphi_{i,j} = 0$$

$$\text{si } \left| \begin{array}{l} i=0 \\ \text{ou } \forall j=0, 1, \dots ny. \\ i=nx \end{array} \right.$$

$$\text{si } \left| \begin{array}{l} j=0 \\ \text{ou } \forall i=0, 1, \dots nx. \\ j=ny \end{array} \right.$$

La valeur discrète des contraintes, en tous ces points, s'obtient en discrétisant les opérateurs qui s'y rattachent.

1° - Contrainte n_1

$$\text{avec } n_1 = \frac{\partial^2 \varphi}{\partial y^2}$$

a) noeuds intérieurs au domaine

au point M_{ij}

$$\left| \begin{array}{l} i=1, \dots nx-1 \\ j=1, \dots ny-1 \end{array} \right. \quad n_1 = \frac{1}{hy^2} (\varphi_{i,j+1} + \varphi_{i,j-1} - 2\varphi_{i,j})$$

b) noeuds appartenant à la frontière

- si $i=0$ ou $i=nx$

$$n_1 = 0 \quad \forall j \text{ car } \varphi_{0,j} = \varphi_{nx,j} = 0$$

- si $j=0$

$$\varphi(x_i, y_{j+1}) = \varphi(x_i, y_j) + hy \varphi'_y(x_i, y_j) + \frac{hy^2}{2!} \varphi''_{y^2}(x_i, y_j)$$

$$j=0 \text{ et } \varphi(x_i, y_0) = \varphi'_y(x_i, y_0) = 0$$

$$n_1 = \frac{2}{hy^2} \varphi_{i,1}$$

- si $j = n_y$

on obtient de façon identique

$$n_1 = \frac{2}{h_y^2} \varphi_{i, n_y-1}$$

2° - Contrainte n_2

$$n_2 = \frac{\partial^2 \varphi}{\partial x^2}$$

a) noeuds intérieurs au domaine

en discrétisant par rapport aux noeuds (i,j) , $(i+1,j)$, $(i-1,j)$ au point M_{ij}

$$\left| \begin{array}{l} i = 1, \dots, n_x-1 \\ j = 1, \dots, n_y-1 \end{array} \right. \implies n_2 = \frac{1}{h_x^2} (\varphi_{i+1,j} + \varphi_{i-1,j} - 2\varphi_{ij})$$

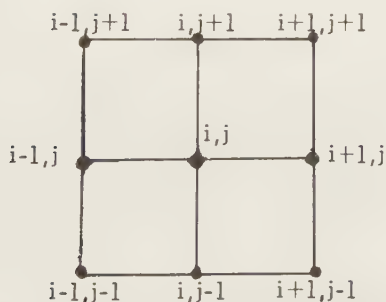
b) noeuds appartenant à la frontière

$$\begin{array}{ll} - \text{ si } j=0 \text{ ou } j=n_y \text{ et } \forall i, & n_2 = 0 \\ - \text{ si } i=0, \forall j=1, \dots, n_y-1 & n_2 = \frac{2}{h_x^2} \varphi_{1,j} \\ - \text{ si } i=n_x, \forall j=1, \dots, n_y-1 & n_2 = \frac{2}{h_x^2} \varphi_{n_x-1,j} \end{array}$$

3° - Contrainte t_3

$$t_3 = - \frac{\partial^2 \varphi}{\partial x \partial y}$$

a) noeuds intérieurs au domaine



en discrétisant par rapport aux noeuds (i,j) , $(i+1,j+1)$, $(i+1,j-1)$, $(i-1,j+1)$, $(i-1,j-1)$ on obtient immédiatement.

$$\begin{array}{l} \forall i = 1, \dots, n_x-1 \\ \forall j = 1, \dots, n_y-1 \end{array} \quad t_3 = - \frac{1}{4.h_x.h_y} (\varphi_{i+1,j+1} + \varphi_{i-1,j-1} - \varphi_{i+1,j-1} - \varphi_{i-1,j+1})$$

b) noeuds appartenant à la frontière

$$\left| \begin{array}{c} j=0 \\ \text{ou} \\ j=n_y \end{array} \right| \quad \forall i=0, \dots, n_x$$

et $t_3 = 0$

$$\left| \begin{array}{c} i=0 \\ \text{ou} \\ i=n_x \end{array} \right| \quad \forall j=0, \dots, n_y$$

$$\text{car } \varphi = \frac{\partial \varphi}{\partial n} = 0 \text{ à la frontière}$$

II - 1. 3. Résultats obtenus

II - 1. 3. 1. Organisation des calculs

Le programme permettant le calcul des contraintes thermiques écrit en langage ALGOL (réf. n° 16) a été exploité sur l'ordinateur IBM 7044 de l'Institut de Mathématiques appliquées de Grenoble.

Trois parties essentielles composaient ce programme

- détermination des valeurs propres extrêmes
- résolution du système linéaire
- calcul des contraintes.

Nous n'insistons pas sur la troisième phase qui ne présentait aucune difficulté.

a) *Détermination des valeurs propres extrêmes.*

- Le calcul de α s'est fait par simple application du théorème de Hadamard.

$$\text{et } \alpha = \left(\frac{4}{h_x^2} + \frac{4}{h_y^2} \right)^2$$

- La méthode itérative avec utilisation des polynômes de Tchebycheff et avec ajustage (voir 1.5.3.2., 3°) s'est montrée être la plus efficace pour le calcul de β . Quelle que soit la dimension n du système, il a en effet suffi d'un petit nombre d'itérations, utilisant un polynôme de degré $p = 6$ pour atteindre un ordre de grandeur de β , très correct, au vu de la résolution qui suivait.

b) *Résolution du système linéaire*

Les dimensions traitées ont été respectivement de

$$n = 176 \quad \left| \begin{array}{c} n_x = 5 \\ n_y = 45 \end{array} \right| \quad \left| \begin{array}{c} h_x = 4 \text{ cm} \\ h_y = 4 \text{ cm} \end{array} \right|$$

$n = 396$	$\left\{ \begin{array}{l} n x = 10 \\ n y = 45 \end{array} \right.$	$\left\{ \begin{array}{l} h x = \underline{2} \text{ cm} \\ h y = 4 \text{ cm} \end{array} \right.$
$n = 836$	$\left\{ \begin{array}{l} n x = 20 \\ n y = 45 \end{array} \right.$	$\left\{ \begin{array}{l} h x = 1 \text{ cm} \\ h y = 4 \text{ cm} \end{array} \right.$
$n = 1691$	$\left\{ \begin{array}{l} n x = 20 \\ n y = 90 \end{array} \right.$	$\left\{ \begin{array}{l} h x = 1 \text{ cm} \\ h y = 2 \text{ cm} \end{array} \right.$

Remarque : Les données numériques du problème qui était posé étaient

$\left\{ \begin{array}{l} OA = 20 \text{ cm} \\ OB = 180 \text{ cm} \end{array} \right.$	(voir figure 2)
$\left\{ \begin{array}{l} E = 8000 \text{ Kg/mm}^2 \\ \alpha = 10^{-5} \text{ par degré} \\ \sigma = 0,3 \end{array} \right.$	

Toutes les résolutions des systèmes se sont bien effectuées, la convergence étant à peu près conforme aux prévisions théoriques.

Exemple : Résolution du système de 1691 équations à 1691 inconnues.

Valeurs propres extrêmes retenues :

$$\left\{ \begin{array}{l} \alpha = 25 \\ \beta = 0,0030 \end{array} \right. \quad (\text{en 15 itérations de degré } p = 6, \text{ la méthode itérative avec utilisation des polynômes de Tchebycheff et avec ajustage donnait } \beta = 0,0034)$$

$$y = \frac{\alpha}{\beta} = 8333$$

5 itérations de degré $p = 200$ ont été effectuées avec un vecteur initial $[X_0] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Le choix de ce vecteur initial a permis d'utiliser une itération de la méthode semi-itérative de Tchebycheff, à la fin de la deuxième itération (voir paragraphe 1.3.5.3. et 1.3.6.).

Le tableau qui suit donne des indications sur le déroulement de la résolution

on lit $\left| \begin{array}{l} \text{en lère colonne :} \end{array} \right.$ le numéro de l'itér $[X_i]$, soit i

en 2ème colonne : la norme vectorielle euclidienne

$$\|X_i - X_{i-1}\| = \sqrt{\sum_{j_i} (x_j^{(i)} - x_j^{(i-1)})^2}$$

en 3ème colonne : $\text{Max}_j |x_j^{(i)} - x_j^{(i-1)}|$ où $x_j^{(i)}$ représente la jième composante de l'itére $[X_i]$

en 4ème colonne :
$$h_{\text{eff}} = \frac{\|X_i - X_{i-1}\|}{\|X_{i-1} - X_{i-2}\|} \quad (\text{n'a de sens que pour } i \geq 2)$$

en 5ème colonne :
$$h_{\text{théor.}} = \frac{1}{T_{200} \left(\frac{\alpha + \beta}{\alpha - \beta} \right)} = h_{200}$$

i	$\ X_i - X_{i-1}\ $	$\text{Max}_j x_j^{(i)} - x_j^{(i-1)} $	h_{eff}	$h_{\text{théor.}}$
1	22225,385	2165,166		0,0250
2	609,997	68,8398	0,02744	"
3	19,124	2,0486	0,03135	"
4	0,304	0,0373	0,01591	"
5	0,0128	0,00186	0,04206	"

La comparaison des colonnes 4 et 5 montre que h_{eff} , s'il n'est pas toujours inférieur ou égal à $h_{\text{théorique}}$, comme cela devrait être, garde cependant une grandeur comparable tout au long de la résolution.

Nous pouvons donc bien préciser la valeur de la solution retenue $[\varphi^*] = [X_5]$ devant la solution exacte du système $[\varphi] = [M]^{-1} [H]$, grâce aux formules d'erreurs introduites en (1.3.4.), et avec $K = 1$ car $[M]$ est symétrique

$$\frac{\|\varphi^* - \varphi\|}{\|\varphi\|} \approx (0,025)^5 \approx 1.10^{-8}$$

$$\|\varphi - \varphi^*\| \approx 0,0128 \frac{0,025}{1-0,025} \approx 0,0003.$$

Cette dernière relation permet de garantir toutes les composantes $> 10^{-3}$ de $[\varphi^*]$. Les plus grandes d'entre elles de l'ordre de 10^4 , ont donc leurs sept chiffres exacts.

II - 1. 3. 2. Résultats obtenus.

Les résultats retenus pour le calcul des contraintes thermiques sont ceux qui correspondent au réseau le plus fin ($n_x = 20, n_y = 90 \implies n = 1691$).

Les contraintes N1, N2 et T3 qui s'en déduisent sont représentées sous forme d'isocontraintes sur les figures V, VI et VII. Elles ne le sont que dans la partie basse de la lingotière ($y \leq 56$ cm), car elles atteignent rapidement des valeurs négligeables, pour des y supérieurs.

La figure 3 donne la répartition initiale des températures. La manière dont avaient été calculées ces températures (de façon discrète) et la pseudo-singularité qu'elles présentent au voisinage du point de contact de la nappe supérieure d'acier liquide à 1500° , avec la paroi de la lingotière ($x = 20$ cm, $y = 41$ cm), rendaient a priori la résolution difficile.

La méthode choisie pour résoudre numériquement l'équation aux dérivées partielles $\Delta \Delta \varphi = g(x, y)$, la méthode des différences finies, étant de plus une méthode de résolution assez grossière, nous n'espérions pas obtenir de résultats très précis.

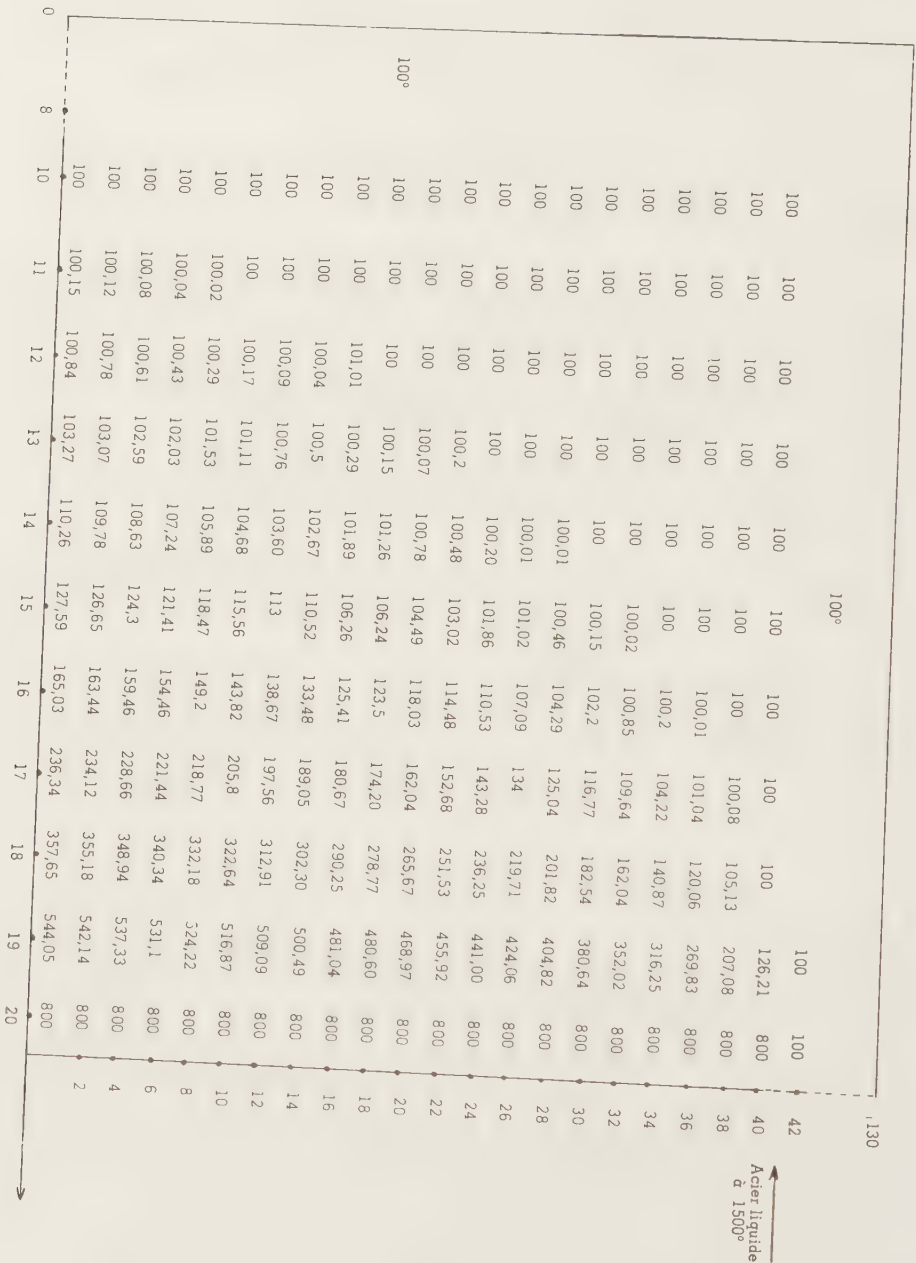
Le tableau qui suit permet de juger de la convergence des composantes du vecteur solution, les φ_{ij} , en des noeuds communs, en fonction du réseau. Il semble que l'ordre de grandeur des valeurs n'est guère perturbé lors du passage de $n = 836$ à $n = 1691$. Le coût de la résolution (3 heures de 7044, programmes écrit en ALGOL, pour le système 1691), et les considérations énoncées ci-dessus, nous ont conduit à ne pas augmenter la dimension du système.

Les résultats que nous présentons n'ont pour ambition que de présenter une physionomie des contraintes thermiques, ce qui n'est pas sans intérêt pour des physiciens.

- Figure 3 -

Répartition initiale des températures (en degré cent.)

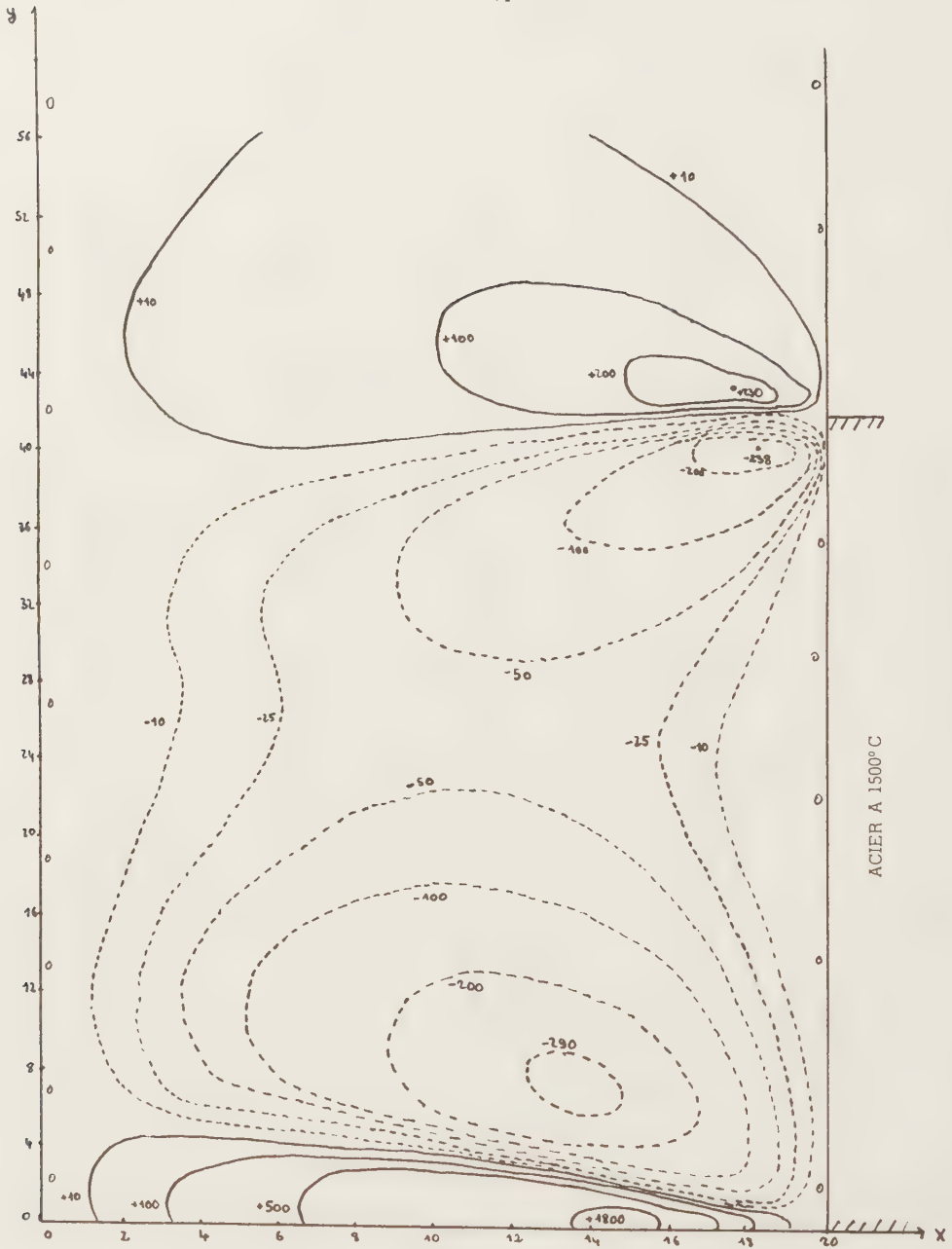
Remarque : Les températures aux noeuds non représentés sont égales à 100° C



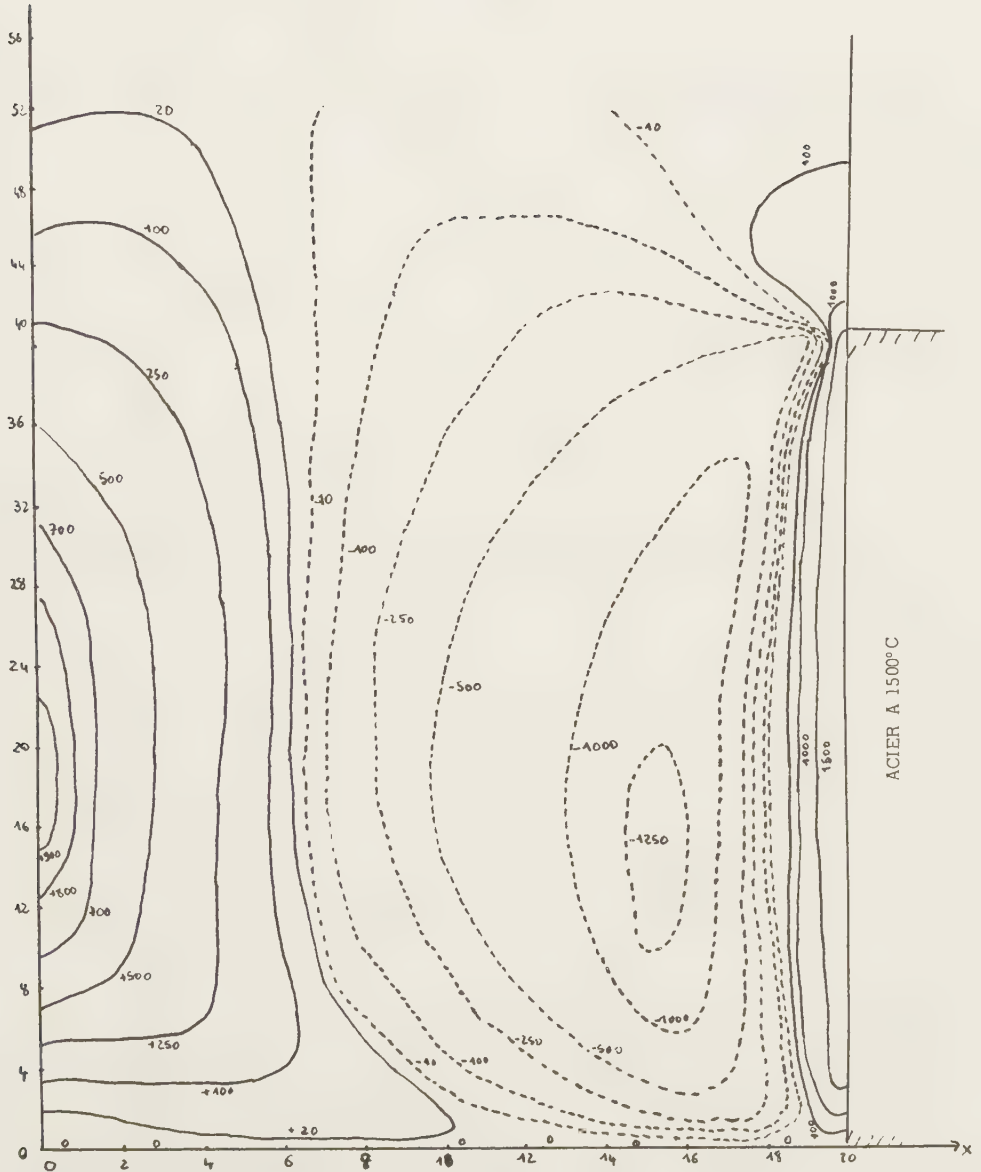
- Valeurs de φ_{ij} pour diverses valeurs de n -

Coordonnées en cms		Valeurs de φ_{ij} en Kg			
Xi	Yj	n = 176	n = 396	n = 836	n = 1691
4	8	4 700,142	4 083,703	3 831,628	3 550,610
	12	6 576,198	5 675,076	5 267,25	5 149,939
	20	6 785,487	6 329,39	5 801,509	5 815,260
	32	2 755,754	4 402,104	4 245,855	4 180,197
	40	671,292	1 947,048	2 316,380	2 149,864
	60	- 18,462	- 22,865	- 14,476	- 20,882
8	8	13 673,821	12 399,796	11 734,903	11 004,002
	12	18 319,286	16 534,713	15 503,812	15 215,978
	20	18 509,03	18 144,97	16 808,491	16 837,79
	32	7 558,338	12 816,216	12 462,082	12 278,75
	40	1 893,918	5 637,7	6 865,727	6 340,64
	60	- 35,332	- 40,774	- 18,530	- 37,218
10	8		16 575,355	15 609,479	14 753,4
	12		21 501,5	20 069,816	19 761,811
	20		23 255,174	21 450,467	21 485,589
	32	(*)	16 623,052	16 064,335	15 844,061
	40		7 158,747	8 896,597	8 158,982
	60		- 42,954	- 17,932	- 39,481
12	8	22 297,464	19 663,104	18 302,221	17 463,575
	12	28 150,463	24 693,409	22 800,015	22 537,49
	20	27 346,494	26 285,384	23 995,656	24 033,35
	32	10 715,206	19 149,253	18 260,591	18 038,692
	40	2 426,835	7 951,581	10 214,716	9 260,139
	60	- 35,466	- 41,237	- 18,446	- 38,57
16	8	22 679,416	17 107,466	14 782,713	14 407,58
	12	26 469,088	20 073,563	17 280,742	17 219,484
	20	24 646,132	21 170,605	18 041,378	18 064,94
	32	8 888,782	17 039,705	15 036,826	14 990,47
	40	1 514,931	6 136,806	9 170,582	7 968,49
	60	- 18,98	- 24,3	- 15,526	- 23,052
18	8		8 802,611	6 698,031	6 589,183
	12		10 087,286	7 657,357	7 649,834
	20		11 001,160	8 210,787	8 219,48
	32	(*)	10 457,325	7 895,180	7 864,21
	40		3 447,097	5 931,96	5 064,72
	60		- 10,045	- 7,081	- 8,89

(*) Ces noeuds ne font pas partie du réseau correspondant à cette valeur de n.

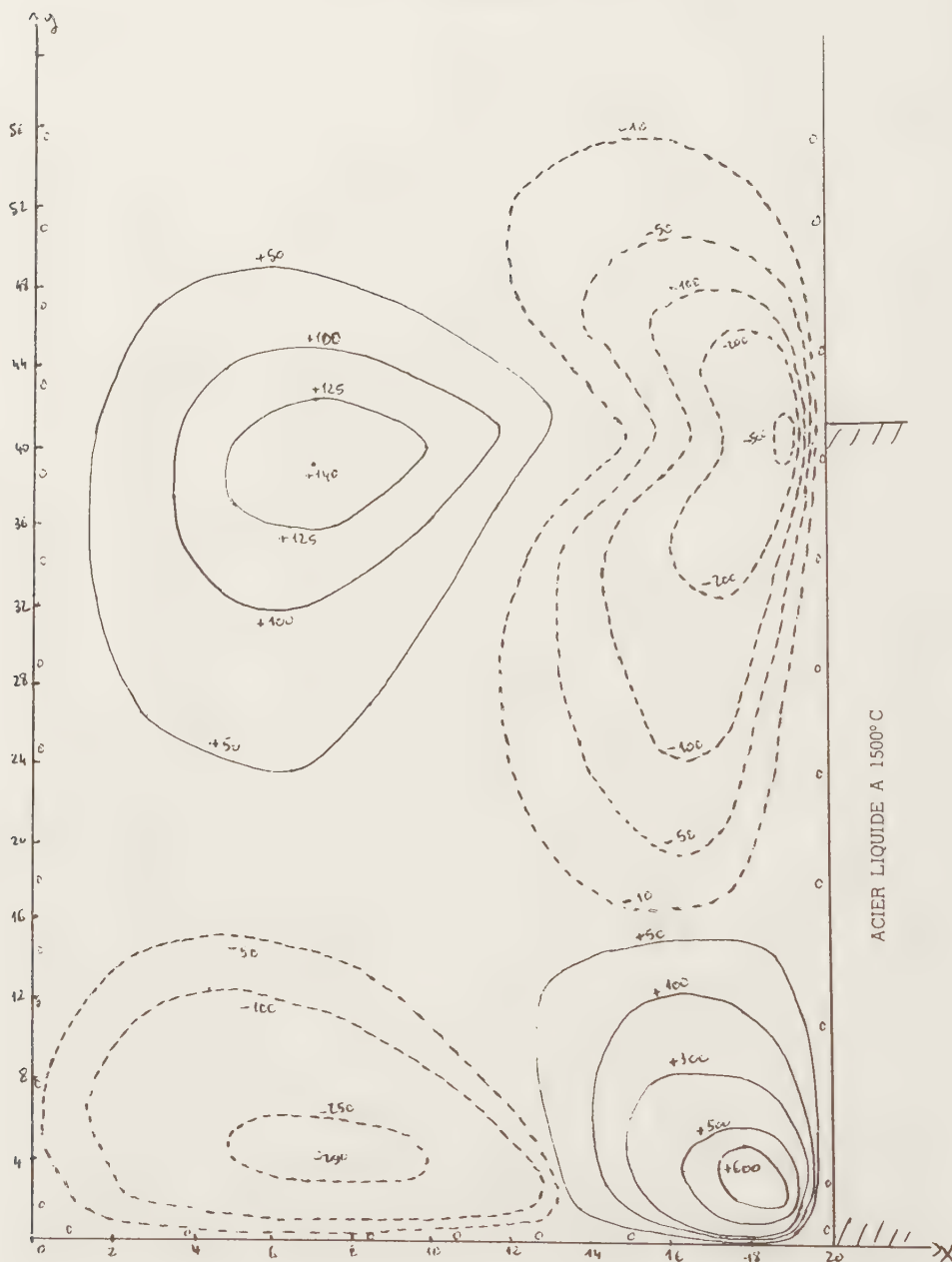
$(y \leq 56)$ 

- Figure VI- Contrainte $N_2 = -\frac{\partial^2 \varphi}{\partial x^2}$ (en Kg/cm²)
($y \leq 56$)



-Figure VII- Contrainte $T_3 = -\frac{\partial^2 \varphi}{\partial x \partial y}$ (en Kg/cm²)

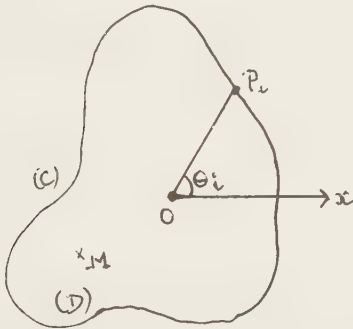
($y \leq 56$)



II - 2. METHODES NUMERIQUES DE RESOLUTION DU PROBLEME DE DIRICHLET

II - 2. 1. Superposition finie de solutions harmoniques.

II - 2. 1. 1. Méthode



Nous recherchons une fonction $\varphi(M)$ définie par

$$\begin{cases} \Delta \varphi(M) = 0 & \forall M \in (D) \\ \varphi(P) = h(P) & \forall P \in (C) \end{cases} \quad (X-2)$$

Définissons une origine O et un axe de référence Ox et repérons le point courant $M \in (D)$ par ses coordonnées polaires

$$\begin{cases} \rho = OM \\ \theta = (\vec{Ox}, \vec{OM}) \end{cases}$$

Nous recherchons $\Psi(M) = \Psi(\rho, \theta)$, solution approchée de (X-2) comme superposition finie des solutions harmoniques élémentaires $\rho^n \cos n \theta$ et $\rho^n \sin n \theta$.

$$\Rightarrow \Psi(M) = \alpha_1 + \sum_{j=1}^p \rho^j (\alpha_{2j} \sin j \theta + \alpha_{2j+1} \cos j \theta) \quad (XI-2)$$

Les N inconnues α_j , en nombre impair ($N=2p+1$) peuvent s'obtenir en interpolant sur (C) la fonction $h(P)$ par le polynôme trigonométrique $\Psi(\rho, \theta) = \Psi(M)$.

Nous fixons N points $P_i \in (C)$, définis par leurs coordonnées (ρ_i, θ_i) et nous imposons en tous ces points $\Psi(\rho_i, \theta_i) = h(P_i) = h(\rho_i, \theta_i)$

$$\Rightarrow \alpha_1 + \sum_{j=1}^p \rho_i^j (\alpha_{2j+1} \cos j \theta_i + \alpha_{2j} \sin j \theta_i) = h_i = h(\rho_i, \theta_i) \quad \forall i=1, \dots, N$$

D'où le système linéaire de N équations à N inconnues (les α_j)

$$[M] [\alpha] = [h] \quad (XII-2)$$

avec $\left| \begin{array}{l} \text{comme éléments de } [M] \end{array} \right.$

$$m_{ij} = \begin{cases} \text{si } j \text{ pair } \rho_i^{j/2} \sin \frac{j}{2} \theta_i \\ \text{si } j \text{ impair } \rho_i^{j-1} \cos \left(\frac{j-1}{2} \right) \theta_i \end{cases} \quad \begin{matrix} (i=1, \dots, N) \\ (j=1, \dots, N) \end{matrix}$$

comme éléments de $[a]$	$a_i \quad i = 1, \dots, N$
comme éléments de $[h]$	$h_i = h(P_i) \quad i = 1, \dots, N.$

Ce système une fois résolu, l'expression (XI-2) permet d'approcher en tout point de (D) la fonction $\Psi(M)$ solution de (X-2).

La précision avec laquelle $\Psi(M)$ approche $\Psi(M)$ dépend de celle avec laquelle $\Psi(P)$ interpole $h(P)$.

En effet, le théorème du maximum, appliqué à la fonction harmonique $\Psi(M) - \Psi(M)$ permet d'affirmer que cette fonction atteint son extrêmuM sur le contour (C)

Donc si	$\forall P \in (C) \quad \Psi(P) - h(P) \leq \varepsilon$	(XIII-2)	(la détermination de ε sera expérimentale).
	$\forall M \in (D) \quad \Psi(M) - \Psi(M) < \varepsilon$		

II - 2. 1. 2. Application à la Recherche Numérique des fonctions de Green.

La fonction de Green attachée à la source $M_0 \in (D)$, et notée $G(M_0, M)$, n'est strictement connue que dans le cas de domaines bien particuliers. Or la connaissance d'une telle fonction peut intervenir dans la résolution d'un certain nombre de problèmes liés au Laplacien (réf. (17), (18)). La détermination de $G(M_0, M)$ peut se faire expérimentalement par l'intermédiaire de cuves rhéoelectriques (réf. (17)). Cette méthode a pour inconvénient de nécessiter l'emploi d'une cuve adaptée au domaine.

Il importe donc de connaître des procédés numériques de recherche de $G(M_0, M)$.

Si nous définissons une fonction $G_1(M_0, M)$ par

$G(M_0, M) = G_1(M_0, M) - \text{Log } M_0, M $
--

la recherche de $G(M_0, M)$, se fera par l'intermédiaire de $G_1(M_0, M)$ qui est solution d'un problème de Dirichlet car

$\Delta G_1(M_0, M) = 0 \quad \forall M \in (D)$ $G_1(M_0, P) = \text{Log } M_0 P \quad \forall P \in (C)$	(XIV - 2)
--	-----------

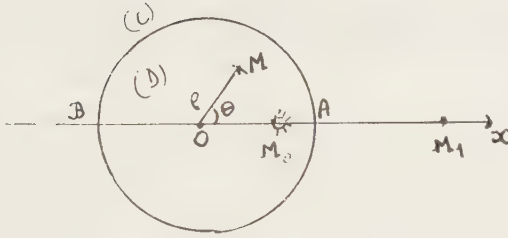
La recherche de $G_1(M_0, M)$ peut se faire par superposition de solutions harmoniques.

1°) Cas d'un domaine circulaire

Ce domaine a été choisi car $G_1(M_0, M)$ y est strictement connu et vaut

$$G_{1_{\text{exact}}}(M_0, M) = \text{Log } |M_1 M_0| + \text{Log } \frac{|OM|}{R}$$

avec $\left\{ \begin{array}{l} R = \text{rayon du cercle (C)} \\ M_1 = \text{inverse de } M_0 \text{ par rapport à (C)} \end{array} \right.$



Le calcul de $G_1(M_0, M)$ a été effectué pour des sources M_0 , se trouvant sur le rayon OA, distantes de l'origine de O, $R/8, 2R/8, \dots, 7R/8$. Le problème étant symétrique, la solution recherchée l'a été sous la forme :

$$\Psi(\rho, \theta) = \sum_{j=1}^N a_j \rho^{(j-1)} \cos(j-1) \theta$$

avec pour points d'interpolation

$$P_i \left\{ \begin{array}{l} \rho_i = R \\ \theta_i = \frac{(i-1)\pi}{(N-1)} \end{array} \right. \quad i = 1, \dots, N$$

⇒ le système linéaire

$$[M] \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} R^{i-1} = [\text{Log } |M_0 P_i|]$$

(XV - 2)

$$\begin{bmatrix} 1 & \cos \theta_1 & \dots & \cos(N-1) \theta_1 \\ 1 & \cos \theta_2 & \dots & \cos(N-1) \theta_2 \\ & & \ddots & \\ 1 & \cos \theta_N & \dots & \cos(N-1) \theta_N \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 R \\ \vdots \\ a_N R^{N-1} \end{bmatrix} = \begin{bmatrix} \text{Log}(M_0 P_1) \\ \text{Log}(M_0 P_2) \\ \vdots \\ \text{Log}(M_0 P_N) \end{bmatrix}$$

Du fait du choix des θ_i , et de la mise en facteur des R^i dans le vecteur solution du système, la matrice $[M]$ obtenue est symétrique donc à valeurs propres réelles. Les essais effectués ayant montré que ces valeurs propres n'étaient pas de même signe, les deux membres du système ont été prémultipliés par $[M]$, pour rendre le système résoluble par la méthode de Richardson.



$$[M]^2 [\alpha^i R^{i-1}] = [M] [\text{Log } |M_0^i P_i|] \quad (\text{XVI} - 2)$$

Les essais ont été faits pour des dimensions < 30 . Le rapport des valeurs propres extrêmes de la matrice $[M]^2$, très faible (≤ 10) a facilité la résolution, qui a toujours donné de très bons résultats en des temps très rapides.

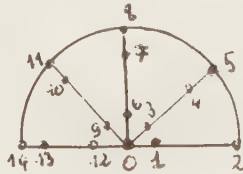
Le tableau qui suit donne, pour quelques sources distinctes M_0 , repérées par $(OM_0 \text{ et } \theta = 0)$ l'erreur maximum d'interpolation déterminée expérimentalement pour diverses valeurs de n , donc la précision avec laquelle $\Psi(\rho, \theta)$ approche $G_{1\text{ex}}(M_0, M)$ en tout point de (D) .

Remarque : $\forall M$, et $M_0 \in (D)$, $G_1(M_0, M) < 5$

OM_0	$n=5$	$n=10$	$n=20$	$n=30$
0	$0,6 \cdot 10^{-7}$	$0,4 \cdot 10^{-6}$	$0,95 \cdot 10^{-6}$	
R/8	$0,12 \cdot 10^{-4}$	$0,4 \cdot 10^{-6}$	$0,83 \cdot 10^{-6}$	$0,9 \cdot 10^{-6}$
5R/8	$0,5 \cdot 10^{-1}$	$0,27 \cdot 10^{-2}$	$0,13 \cdot 10^{-4}$	$0,11 \cdot 10^{-5}$
6R/8	0,165	$0,2 \cdot 10^{-1}$	$0,63 \cdot 10^{-3}$	$0,25 \cdot 10^{-4}$
7R/8	0,508	0,155	$0,22 \cdot 10^{-1}$	$0,43 \cdot 10^{-2}$

Pour $n=30$ et dans le cas le plus défavorable, c'est-à-dire lorsque la source est proche de la frontière ($OM=7R/8$), Ψ et G_1 ont les trois premiers chiffres significatifs exacts. Des résultats plus valables auraient peut être pu être obtenus en augmentant le nombre d'inconnues.

Quelques résultats



$OM_0 = 5R/8$ ($n=30$)

$OM_0 = 7R/8$ ($n=30$)

Point n°	$G_{1\text{ex}}$	$G_{1\text{app.}} = \Psi$
0	3,6763007	3,6762992
1	3,5064016	3,5064017
2	2,6954714	2,6954715
3	3,5668763	3,5668764
4	3,3835386	3,3835391
5	3,3364241	3,3364235
6	3,6883611	3,6883612
7	3,7756140	3,7756144
8	3,8411773	3,8411771
9	3,7860230	3,7860232
10	3,9926380	3,9926381
11	4,0871826	4,0871829
12	3,8214827	3,8214828
13	4,0607124	4,0607123
14	4,1618085	4,1618081

Point n°	$G_{1\text{ex}}$	$G_{1\text{app.}} = \Psi$
0	3,6763007	3,6762917
1	3,4294406	3,4294281
2	1,5968591	1,5968586
3	3,5247284	3,5247186
4	3,3323068	3,3352999
5	3,3571493	3,3583168
6	3,6996716	3,6996652
7	3,8553700	3,8555411
8	3,9605530	3,9598455
9	3,8290165	3,8290117
10	4,1053648	4,1052822
11	4,2261168	4,2263254
12	3,8741264	3,8741221
13	4,1808567	4,1800709
14	4,3049093	4,3049090

2°) Remarques - extension à un domaine quelconque

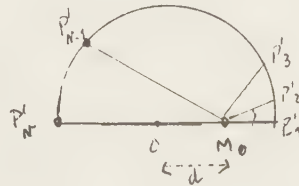
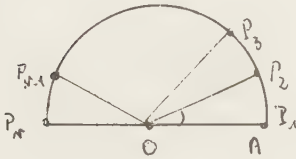
Les calculs dont il est fait mention ci-dessus montraient que les résultats obtenus étaient d'autant moins bons que la source était proche de la frontière.

De plus, l'exemple choisi n'avait pour intérêt que de tester la méthode de résolution, avant de l'appliquer à d'autres domaines.

A cet effet, il a semblé intéressant de décentrer l'origine de référence en la confondant avec la source M_0 , distante au centre du cercle de d , et en choisissant de nouveaux points l'interpolation P_i' distincts des P_i .

$$P_i \left\{ \begin{array}{l} \text{origine } O \\ \left\{ \begin{array}{l} \rho_i = OP_i = R \\ \theta_i = (\vec{OP}_1, \vec{OP}_i) = \frac{(i-1)\pi}{N-1} \end{array} \right. \end{array} \right.$$

$$P_i' \left\{ \begin{array}{l} \text{origine } M_0 \\ \left\{ \begin{array}{l} \rho_i' = M_0 P_i' = f(R, d, \theta_i) \\ \theta_i' = \frac{(i-1)\pi}{(N-1)} = \theta_i \end{array} \right. \end{array} \right.$$



Les N nouveaux points d'interpolation P_i' , en plus grand nombre à proximité de la source M_0 , devaient permettre une meilleure interpolation dans ce voisinage.

De plus, l'utilisation de l'excentricité ainsi introduite devait renseigner sur la validité de la méthode de superposition de solutions harmoniques lorsqu'on l'applique à des domaines autres que le cercle.

La matrice $[M']$ du système à résoudre est distincte de la matrice $[M]$ définie en (XV-2), en ce sens que des termes en ρ_i^j interviennent dans sa construction alors qu'auparavant, valant R^j , ils étaient mis en facteur

$$\Rightarrow \boxed{[M'] [\alpha] = [\text{Log}|M_0 P_i'|]} \quad (\text{XVII} - 2)$$

$$\text{avec } [M'] = \begin{bmatrix} 1 & \rho_1' & \cos \theta_1' & \dots & \rho_1'^{N-1} & \cos ((N-1) \theta_1') \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \rho_i' & \cos \theta_i' & \dots & \rho_i'^{N-1} & \cos ((N-1) \theta_i') \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \rho_N' & \cos \theta_N' & \dots & \rho_N'^{N-1} & \cos ((N-1) \theta_N') \end{bmatrix}$$

Cette matrice n'est plus symétrique, donc ses valeurs propres ne sont pas obligatoirement réelles.

Le système à résoudre est alors

$$[M']^t [M'] [\underline{a}_1] = [M']^t [Log \cdot [M_0 P_i]] \quad (XVIII - 2)$$

Les premiers essais effectués pour $N = 5$, $D = R/2$ et $R = 39,5$ ont été aussitôt interrompus car la plus grande valeur propre $\propto$ était de l'ordre de 10^{11} et la plus petite était impossible à calculer (l'ordinateur utilisé, la C.A.E. 510 travaille avec 8 chiffres significatifs).

Ce phénomène était dû à la présence des termes en $\rho_i^{\prime j}$ dans la matrice de résolution $[M']^t [M]$.

Deux suppositions pouvaient expliquer ce trop grand rapport $\frac{\lambda_{\max}}{\lambda_{\min}}$

- l'importance en grandeur des ρ_i'
- une trop grande disproportion entre le plus grand des ρ_i' , soit ici $\rho_N' = R + d$ et le plus petit ici $R - d = \rho_1'$.

Remarque : cas de $n = 2$

$\rho_1' = R - d$	$\rho_2' = R + d$
$\theta_1' = 0$	$\theta_2' = \pi$

$$\Rightarrow [M'] = \begin{bmatrix} 1 & R-d \\ 1 & -(R+d) \end{bmatrix}$$

$$\Rightarrow \quad [M']^t [M'] = \begin{bmatrix} 2 & -2d \\ -2d & +2(R^2 + d^2) \end{bmatrix}$$

On trouve immédiatement dans ce cas simple

$$y_0 = \frac{\lambda_{\max}}{\lambda_{\min}} = \left(\frac{(1+R^2+d^2) + \sqrt{(1+R^2+d^2)^2 - 4R^2}}{2R} \right)^2$$

- à d constant y_0 est minimum pour $R=1$ et croît très rapidement dès que R s'éloigne de cette valeur
- à R constant y_0 est fonction croissante de d , se comportant comme $d^2 \implies y_0$ minimum pour $d=0$.

a) Influence de l'unité de référence

Nous nous sommes tout d'abord attachés à analyser l'influence de l'introduction de termes en ρ_i dans la matrice de résolution en nous ramenant à l'étude du problème simplifié, où tous les ρ_i sont constants et valent R (M_0 en 0 et $d=0$).

$$[M] = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & R \cos \theta_2 & \dots & R^{N-1} \cos(N-1)\theta_2 \\ \dots & \dots & \dots & \dots \\ 1 & R \cos \theta_N & \dots & R^{N-1} \cos(N-1)\theta_N \end{bmatrix} \theta_i = \frac{(i-1) \pi}{(N-1)}$$

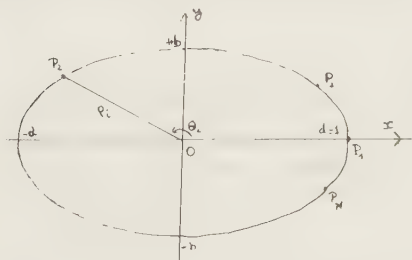
Nous avons calculé pour deux dimensions ($N=5$ et $N=10$), les valeurs propres extrêmes $\alpha = \lambda \max$ et $\beta = \lambda \min$ de la matrice $[M']^t [M']$, en faisant varier R de 0.1 à 5. Les résultats obtenus sont groupés ci-dessous

R	N 5			N 10		
	α	β	α/β	α	β	α/β
5	1953250	63,56	30730	$0,4 \cdot 10^{14}$	$0,5 \cdot 10^5$	$0,8 \cdot 10^8$
2	1283	5,13	250,3	$0,26 \cdot 10^7$	276	9503
1,5	130	1,48	29	13862	14	1405
1,25	30,71	4,204	7,305	560	4,67	120
1,1	11,633	3,402	3,42			
1	6,56	2,44	2,69	11,386	4,52	2,52
0,9	5,386	1,505	3,58			
0,75	5,101	0,435	11,715			
0,5	5,014	0,0182	274,43	10,007	0,033	305
0,1	5	0,0003	18067	10,000	0,00059	18779

Une petite variation sur R influe beaucoup sur le rapport α/β . Minimum pour $R = 1$, il prend, dès qu'il s'en éloigne par valeur supérieure, ou par valeur inférieure, une grandeur telle que le système devient difficilement résoluble par méthode de Richardson. Ce rapport $\frac{\alpha}{\beta}$ semble, de plus, être fonction croissante de n .

b) Influence de l'excentricité

Afin d'étudier l'influence de l'excentricité sur le rapport α/β , donc sur la possibilité de résolution par la méthode de Richardson de ce type de problèmes, nous avons considéré comme domaine (D) l'ellipse de grand axe $a = 1$ et de petit axe b . Et pour rendre cette étude plus générale nous ne nous sommes plus limités au problème symétrique et avons introduit des termes en $\rho_i^{\frac{1}{2}} \sin i \theta_j$



La matrice $[M]$, définie en (XII-2), est construite à partir des points d'interpolation

$$\theta_i = (i-1) \frac{2\pi}{N}$$

$$\rho_i = \frac{ab}{\sqrt{b^2 \cos^2 \theta_i + a^2 \sin^2 \theta_i}}$$

Après avoir vérifié que les valeurs propres de la matrice $[M]$ n'étaient pas de même signe, nous avons calculé pour diverses valeurs de $\frac{b}{a} = b$ et pour deux valeurs de $N = 5$ et 15 , les valeurs propres extrêmes α et β de la matrice $[M]^t [M]$

a	b	α / β	
		$N = 5$	$N = 15$
1	1	≈ 2	≈ 2
1	0,8	$\approx 3,4$	$\approx 24,60$
1	0,5	≈ 10	$\approx 33,4$
1	0,2	≈ 74	≈ 50
1	0,1	≈ 331	≈ 250
1	0,01	≈ 3700	≈ 3250

Si le rôle de N ne semble pas être ici important, l'excentricité influe sur le rapport α / β dans le sens prévu. α / β , minimum pour le cercle, croît lorsque l'ellipse s'applatit. Ce phénomène est moins marqué cependant, que celui qui était dû à la présence de termes en $\rho_i > 1$.

Remarque : quelques essais supplémentaires ont été faits pour l'ellipse avec cette fois des valeurs du grand axe $a > 1$.

Pour	$\alpha = 5$	$b/a = 1$	$n = 5$	$\implies$	$\alpha/\beta \approx 25$
	$\alpha = 5$	$b/a = 0,8$	$n = 5$	$\implies$	$\alpha/\beta \approx 220$
	$\alpha = 5$	$b/a = 0,5$	$n = 5$	$\implies$	$\alpha/\beta \approx 350$
	$\alpha = 0,5$	$b/a = 1$	$n = 5$	$\implies$	$\alpha/\beta \approx 8$
	$\alpha = 0,5$	$b/a = 0,1$	$n = 5$	$\implies$	$\alpha/\beta \approx 3200$
	$\alpha = 0,5$	$b/a = 0,01$	$n = 5$	$\implies$	$\alpha/\beta \approx 37800$
	$\alpha = 2$	$b/a = 1$	$n = 15$	$\implies$	$\alpha/\beta \approx 4096$
	$\alpha = 2$	$b/a = 0,1$	$n = 15$	$\implies$	$\alpha/\beta \approx 23693$

c) Conclusion

Les essais n'ont pas été poursuivis pour d'autres domaines. Ils mériteraient de l'être.

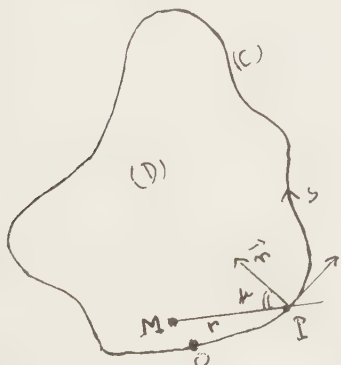
Cette méthode de résolution du problème de Dirichlet semble être, lorsqu'elle est utilisable, intéressante par la simplicité de mise en équation et par la précision qu'elle fournit.

Les résultats expérimentaux, présentés ci-dessus, montrent cependant, que la validité et l'intérêt de la méthode dépendent du domaine. Un domaine très allongé ne rendra pas aisée cette résolution. De toutes façons, pour rendre l'utilisation de la méthode de Richardson profitable pour résoudre le système linéaire qui s'introduit, des précautions nécessaires devront toujours être prises :

- 1) par une "mise en facteur appropriée", il faudra rendre le plus grand des ρ_i , qui intervient dans la matrice $[M]$, égal à 1.
- 2) ne pas oublier de prémultiplier les deux membres s de l'égalité par la matrice transposée $[M]^t$.

II - 2. 2. Utilisation de la méthode du potentiel

II - 2. 2. 1. Principes généraux et méthode de résolution



Supposons continuellement réparti sur (C) , un potentiel de double couche de densité $g(s)$, (C) étant régulier.

Si P (d'abscisse curviligne s), est le point courant de (C) , le potentiel induit en $M \in (D)$ vaut par définition

$$\varphi(M) = \frac{1}{\pi} \int_C g(s) \frac{\cos \mu(M, P)}{r(M, P)} ds \quad (\text{XIX-2})$$

avec $\begin{cases} \mu(M, P) = \text{angle formé par } \overrightarrow{PM} \text{ et la} \\ \text{normale } \vec{n} \text{ à } (C), \text{ orientée vers l'in-} \\ \text{térieur en } P. \\ r(M, P) = \text{distance des deux points} \\ M \text{ et } P. \end{cases}$

De plus, si M tend vers un point P de la frontière, on sait que

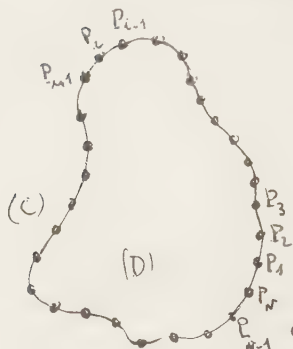
$$\varphi(P_0) = g(P_0) + \frac{1}{\pi} \int_C g(s) \frac{\cos \mu(P_0, P)}{r(P_0, P)} ds$$

Par définition du problème de Dirichlet posé (X-2), $\varphi(P) = h(P)$

$$\Rightarrow \quad h(P_0) = g(P_0) + \frac{1}{\pi} \int_C g(s) \frac{\cos \mu(P_0, P)}{r(P_0, P)} ds \quad (\text{XX-2})$$

La densité $g(s)$ n'est pratiquement jamais connue sous forme mathématique. Le calcul du potentiel au point M ne pourra donc se faire que numériquement à l'aide de la fonction $g(s)$, que nous devons déterminer de façon discrète en la calculant en des points P_i de (C) .

1° - Calcul de $g(s)$ sous forme discrète



Partitionnons la frontière (C) en n sections élémentaires Δ_i limitées par les points P_i , $i = 1, \dots, N$.

$$(C) = \bigcup_{i=1}^N \Delta_i \quad \text{avec} \quad \begin{cases} \Delta_i = \widehat{P_i P_{i+1}} & i=1, \dots, N-1 \\ \Delta_N = \widehat{P_N P_1} \end{cases}$$

En chacun de ces points P_i , appliquons l'égalité (XX-2) :

$$g(P_i) + \frac{1}{\pi} \int_{(C)} g(s) \frac{\cos \mu(P_i, P)}{r(P_i, P)} ds = h(P_i)$$

Par un procédé de quadrature (réf. [9]), l'intégrale $\frac{1}{\pi} \int_{(C)} g(s) \frac{\cos \mu}{r} ds$ peut être approchée par

$$\sum_{j=1}^N \alpha_{ij} g(s_j)$$

avec $g(s_j) = g(P_j) = g_j$, les points P_j étant ceux définis ci-dessus.

$$\boxed{g_i + \sum_{j=1}^N \alpha_{ij} g_j = h_i} \quad , \text{ si } h_i = h(P_i)$$

Dans cette relation $h_i = h(P_i)$ est connu ainsi que les α_{ij} qui dépendent de la technique de quadrature choisie et du contour.

Le système à résoudre s'obtient en écrivant que cette relation est vérifiée aux N points choisis P_i .

$$\Rightarrow \boxed{[A] [G] = [H]} \quad (XXI - 2)$$

2° - Calcul de $\varphi(M)$ en un point quelconque de (D)

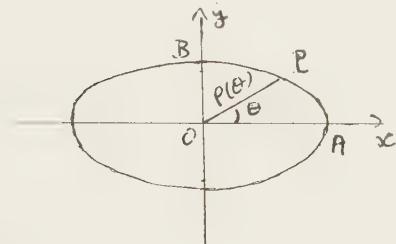
Le système (XXI-2) résolu donne la répartition discrète de la densité $g(s)$ aux points $P_i \in (C)$.

Le potentiel en M s'obtient en calculant l'intégrale (XIX - 2) par quadrature numérique, en se référant aux mêmes points P_i .

II - 2. 2. Applications numériques

1° - Résolution du problème de Dirichlet pour une ellipse

a) Calcul de la densité



Soit l'ellipse de grand axe $OA = a$, de petit axe $OB = b = OA \cdot c = a \cdot c$. Un point P du contour, sera repéré par ses coordonnées polaires (origine O , axe de référence $O\vec{x}$).

$$P = \begin{cases} \theta \\ \rho(\theta) = \frac{ac}{\sqrt{(c^2 \cos^2 \theta + \sin^2 \theta)}} \end{cases}$$

introduisons N points P_i définis par
$$\begin{cases} \theta_i = \frac{2\pi(i-1)}{N} \\ \rho(\theta_i) \end{cases} \quad i = 1, \dots, N$$

La densité $g(s)$ aux points P_i , que nous notons g_i s'obtient en discrétisant l'intégrale de

$$h_i = g_i + \frac{1}{\pi} \int_C \frac{\cos \mu(P_i, P)}{r(P_i, P)} g(s) ds \quad (\text{XXII-2})$$

$$\text{si } h_i = h(P_i)$$

L'intégrale curviligne pourra être mise sous la forme, $f_i(\theta)$ et $g(\theta)$ étant périodiques :

$$\int_0^{2\pi} f_i(\theta) g(\theta) d\theta.$$

Si la méthode composite est adoptée (réf. [9]), cette intégrale périodique vaudra numériquement

$$\frac{2\pi}{n} \sum_{j=1}^N f_i(\theta_j) g_j$$

$$\text{or } f_i(\theta) = \left(\frac{\cos \mu(P_i, P)}{r(P_i, P)} \right) \cdot \left(\frac{ds}{d\theta} \right)$$

$$\text{avec } \frac{ds}{d\theta} = \frac{ca \sqrt{(c^2 \cos^2 \theta + \sin^2 \theta)^2 + (c^2 - 1)^2 \sin^2 \theta \cos^2 \theta}}{(c^2 \cos^2 \theta + \sin^2 \theta)^{3/2}}$$

$$\frac{\cos \mu(P_i, P)}{r(P_i, P)} = \frac{1 - \rho(\theta) \rho_i(\theta_i) \left[\frac{\cos \theta_1 \cos \theta}{\alpha^2} - \frac{\sin \theta_1 \sin \theta}{c^2 \alpha^2} \right]}{r^2(\theta, \theta_i) \sqrt{\frac{\rho^2(\theta) \cos^2 \theta}{\alpha^4} + \frac{\rho^2(\theta) \sin^2 \theta}{c^4 \alpha^4}}}$$

$$\text{D'où } \left\{ \begin{array}{l} f_i(\theta_j) = \left(\frac{ds}{d\theta} \right)_{\theta=\theta_j} \frac{\cos \mu(P_i, P_j)}{r(P_i, P_j)} \quad \text{si } j \neq i \\ f_i(\theta_i) = \left(\frac{ds}{d\theta} \right)_{\theta=\theta_i} \lim_{P_j \rightarrow P_i} \left(\frac{\cos \mu(P_i, P_j)}{r(P_i, P_j)} \right) \quad \text{si } j = i \end{array} \right.$$

$$\text{or } \lim_{P_j \rightarrow P_i} \left(\frac{\cos \mu(P_i, P_j)}{r(P_i, P_j)} \right) = \frac{1}{2\mathcal{R}_{(P_i)}} \quad \mathcal{R}_{(P_i)} \text{ étant le rayon de courbure au point } P_i$$

$$= \frac{c}{2\alpha} \left(\frac{c^2 \cos^2 \theta_i + \sin^2 \theta_i}{\sin^2 \theta_i + c^4 \cos^2 \theta_i} \right)^{3/2}$$

en posant $\underline{f_i(\theta_j) = f_{ij}}$

la relation (XXII-2) devient

$$\boxed{h_i = g_i \left(1 + \frac{2}{N} f_{1i} \right) + \frac{2}{N} \sum_{\substack{j=1 \\ j \neq i}}^N f_{ij} g_j} \quad (\text{XXIII} - 2)$$

L'écriture des n relations ci-dessus avec $i = 1, \dots, n$ conduit au système

$$[A] [g] = [h]$$

avec $[A] = \begin{bmatrix} 1 & 0 & \dots & 0 \\ & 1 & 0 & 0 \\ & & \ddots & \\ & 0 & 1 & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} + \frac{2}{N} \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \dots & f_{nn} \end{bmatrix}$

b) Calcul du potentiel en un point quelconque intérieur à l'ellipse

Ce calcul s'effectue avec

$$\varphi(M) = \frac{1}{\pi} \int_C g(s) \frac{\cos \mu(M, P)}{r(M, P)} ds$$

La même formule de quadrature, utilisant les mêmes points P_j donne immédiatement :

$$\varphi(M) = \varphi(\rho, \theta) = \frac{1}{\pi} \frac{2\pi}{N} \sum_{j=1}^N f_j g_j = \frac{2}{N} \sum_{j=1}^N f_j g_j \quad (\text{XXIV} - 2)$$

avec $f_j = \frac{\cos \mu(M, P_j)}{r(M, P_j)} \left(\frac{ds}{d\theta} \right)_{\theta=\theta_j}$, qui est analogue au f_{ij} , défini ci-dessus, en remplaçant θ_i par $\theta, \rho(\theta_i)$

par ρ
si $M = \begin{cases} \theta \\ \rho \end{cases}$

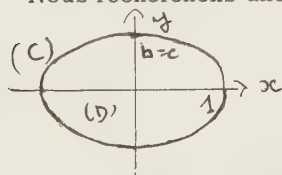
c) Résultats numériques

Remarque préliminaire : Une étude des éléments f_{ij} de la matrice $[A]$ montre que les dimensions de l'ellipse n'y interviennent que dans le rapport $\frac{b}{a} = c$, rapport du petit axe au grand axe. Le choix des unités de référence ne joue donc ici aucun rôle dans la matrice du système, qui ne dépend que de la forme du domaine, de la répartition des points P_i (définis par leur angle θ_i), et de la technique de quadrature utilisée. Nous ne rencontrons plus ici le phénomène, du à l'unité de référence choisie, qui intervenait dans la résolution du problème de Dirichlet, par superposition de solutions harmoniques.

Pour cette raison, les résultats, présentés ci-dessous sont relatifs à des ellipse dont le grand axe vaut toujours $1 = a$.

C-1 Problèmes résolus

Nous recherchons une fonction $\varphi(x, y) = \varphi(M)$ définie par



$$\begin{cases} \Delta \varphi = 0 & \text{si } M \in (D) \\ \varphi(P) = e^y \cos x & \text{si } P \left\{ \begin{matrix} x \\ y \end{matrix} \right\} \in (C) \end{cases}$$

$\varphi(x, y)$ vaut donc $\forall M(x, y) \in (D)$

$\varphi_{ex} = e^y \cos x$, ce qui permet de tester la méthode.

La résolution numérique a été faite pour deux ellipses définies par $c=1$ (cercle de rayon 1) et $c=0,5$ (ellipse d'excentricité $e=0,86$), pour des dimensions $n=10$, $n=20$ et $n=50$.

C-2 Cas du domaine circulaire

* - Calcul de la densité $g(s)$ aux points $P_i \in (C)$, définis par

$$\text{leurs coordonnées} \quad \begin{cases} \theta_i = \frac{2(i-1)\pi}{n} \\ \rho_i = 1 \end{cases}$$

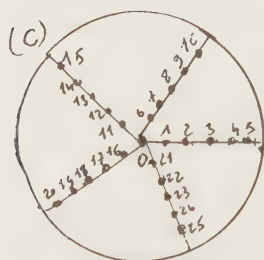
Quelle que soit la dimension n , donc le nombre d'inconnues g_i , les valeurs propres de la matrice $[A]$, toutes positives, étaient groupées dans l'intervalle $[\beta = 1, \alpha = 2]$. La résolution du système $[A][G] = [H]$ par la méthode de Richardson avec balayage, a donc été très facilitée, et les résultats obtenus en des temps très rapides étaient très précis : ainsi le vecteur solution $[G^*]$ était tel que $[A][G^*] - [H]$ avait ses composantes toutes nulles pour $n = 50$.

Le tableau ci-après donne pour les points P_i , communs aux trois résolutions, la densité g_i . On y vérifie que la symétrie par rapport à Oy , due au choix de la fonction $\varphi_{ex} = \cos x e^y$ est respectée.

P_i		g_i		
ρ_i	θ_i	$n = 10$	$n = 20$	$n = 50$
1	0	0,040302433	0,040302312	0,040302306
1	$\pi/5$	0,74237651	0,74237637	0,74237639
1	$2\pi/5$	1,9658363	1,9658361	1,9658362
1	$3\pi/5$	1,9658364	1,9658362	1,9658362
1	$4\pi/5$	0,74237657	0,74237642	0,74237644
1	π	0,040302450	0,040302335	0,040302325
1	$6\pi/5$	-0,11654929	-0,11654942	-0,11654942
1	$7\pi/5$	-0,13196666	-0,13196678	-0,13196681
1	$8\pi/5$	-0,13196667	-0,13196678	-0,13196679
1	$9\pi/5$	-0,11654930	-0,11654942	-0,11654943

* - Calcul du potentiel $\varphi(M)$ en un point intérieur à (D)

Le calcul du potentiel a été fait, à partir des g_k calculés dans la première phase, pour les points M de (D), numérotés sur la figure ci-contre. Il a tout d'abord été effectué, conformément à (XXIV-2) c'est-à-dire :

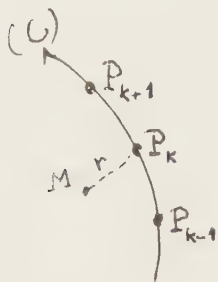


$$\varphi^*(M) = \frac{2}{n} \sum_{j=1}^n f_j g_j \quad (\text{XXIVbis} - 2)$$

$$\text{avec} \quad f_j = \frac{\cos \mu(M, P_j)}{r(M, P_j)} \left(\frac{ds}{d\theta} \right)_{\theta = \theta_j}$$

Les valeurs ainsi obtenues pour $n = 50$, se trouvent dans le tableau [1], en 3^{ième} colonne, le résultat exact $\varphi(M) = \cos x e^y$, s'y trouvant en deuxième colonne. On constate que les résultats, excellents pour les points centraux, le

sont moins dès que le point M se rapproche du contour. Une ébauche d'explication peut être faite, le calcul de la densité ne pouvant être mis en cause.



Supposons que le point du contour P_k (auquel est attachée la densité ponctuelle g_k), soit parmi tous les P_i ($i=1, \dots, n$), le plus proche de M. Dans le calcul de $\varphi^*(M)$, la valeur g_k intervient, multipliée par le facteur que nous avons appelé f_k . Si l'on regarde l'expression exacte de f_k (voir (a)), on constate que $r(M, P_k) = MP_k$ y intervient au carré, au dénominateur. f_k est donc numériquement mal défini, et ce d'autant plus que le numérateur de f_k est petit. L'expérience montre que le maniement de telles quantités peut conduire à des résultats effectifs parfois éloignés de ce qu'ils devraient être. Une amélioration

du calcul du potentiel peut être faite en modifiant l'intégrale à calculer, en ce point.

On sait démontrer que $\int_C \frac{\cos \mu(M, P)}{r(M, P)} ds = 2\pi$ pour $M \in (D)$

$$\text{or } \varphi(M) = \frac{1}{\pi} \int_C \cos \frac{\mu}{r} g(s) ds = \frac{1}{\pi} \int_C \frac{\cos \mu}{r} [g(s) - g(s_k) + g(s_k)] ds$$

$$\varphi(M) = 2 g(s_k) + \frac{1}{\pi} \int_C \frac{\cos \mu}{r} (g(s) - g(s_k)) ds$$

Le facteur f_k disparaît alors, lorsqu'on applique à cette intégrale la même technique de quadrature et,

$$\varphi^{**}(M) = 2 g_k + \sum_{\substack{j=1 \\ j \neq k}}^n f_j (g_j - g_k) \quad (\text{XXV} - 2)$$

On constate en consultant la quatrième colonne du tableau 1, que la valeur $\varphi^{**}(M)$ ainsi obtenue est un peu meilleure, excepté au point 5, pour les points limitrophes. Si l'amélioration n'est ici guère sensible, il n'en sera pas de même, pour les résultats correspondant à l'ellipse, résultats décrits par ailleurs, et qui justifient cette transformation de l'intégrale. (- voir tableau 1 page suivante -)

C - 5 Cas de l'ellipse ($a=1$, $b=0,5$, excentricité $e=0,86$)

* Calcul de la densité $g(s)$

La discrétisation de la densité $g(s)$ sur le contour de l'ellipse (C) s'est encore faite sur 10 points, 20 points et 50 points.

Point n°	φ_{exact}	φ^*	φ^{**}
0	1	0,99999998	0,99999998
1	0,98614323	0,98614325	0,98614322
2	0,94495695	0,94495696	0,94495697
3	0,87758256	0,87758256	0,87758259
4	0,78588726	0,78588723	0,78588729
5	0,67241224	0,67241607	0,67208641
6	1,1702092	1,1702092	1,1702092
7	1,3657508	1,3657508	1,3657508
8	1,5896979	1,5896979	1,5896979
9	1,8453437	1,8453436	1,8453436
10	2,1361759	2,1366214	2,1363496
11	1,0929125	1,0929125	1,0929125
12	1,1724756	1,1724756	1,1724755
13	1,2333641	1,2333641	1,2333641
14	1,2696720	1,2696720	1,2696720
15	1,2749676	1,2751235	1,2751424
16	0,89845174	0,89845177	0,89845174
17	0,79236000	0,79235999	0,79235998
18	0,68520324	0,68520322	0,68520319
19	0,57986785	0,57986782	0,57986785
20	0,47868079	0,47865413	0,47868314
21	0,85228331	0,85228331	0,85228327
22	0,72445666	0,72445666	0,72445665
23	0,61415220	0,61415216	0,61415221
24	0,51922922	0,51922921	0,51922920
25	0,43776329	0,43773420	0,43774701

Tableau 1

Les valeurs propres, pour ces trois calculs, sont toujours restées groupées dans l'intervalle ($\beta=0,75$, $\alpha=2,5$). La résolution par Richardson s'est donc encore révélée être très satisfaisante.

Le tableau n° 2 donne les densités g_i aux points P_i , communs aux trois réseaux.

$$\text{avec } P_i = \begin{cases} \theta_i \\ \rho_i = 0,5 \\ \sqrt{(0,25 \cos^2 \theta_i + \sin^2 \theta_i)} \end{cases}$$

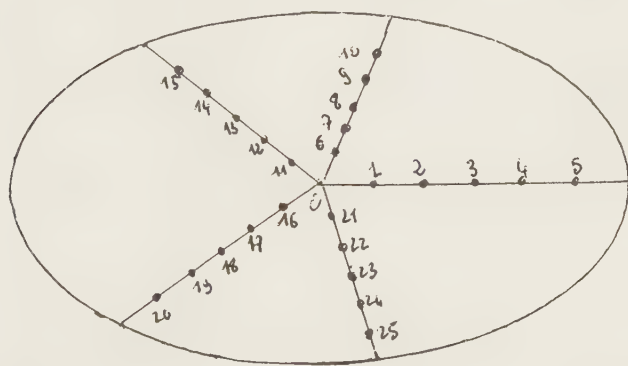
(- voir tableau 2 page suivante -)

P_i	q_i		
θ_i	$n = 10$	$n = 20$	$n = 50$
0	0,16401070	0,15906174	0,15893394
$\pi/5$	1,0359065	1,0392136	1,0392159
$2\pi/5$	1,4018919	1,4051107	1,4051065
$3\pi/5$	1,4018919	1,4051107	1,4051066
$4\pi/5$	1,0359065	1,0392136	1,0392159
π	0,16401072	0,15906177	0,15893396
$6\pi/5$	-0,038593676	-0,048819460	-0,048783064
$7\pi/5$	-0,055898149	-0,062444435	-0,062424589
$8\pi/5$	-0,055898135	-0,062444428	-0,062424575
$9\pi/5$	-0,038593683	-0,048819480	-0,048783083

Tableau 2

* Calcul du potentiel $\varphi(M)$ en divers points de (D)

Les valeurs approchées de $\varphi(M)$, le calcul de l'intégrale s'étant effectué des deux façons décrites ci-dessus (XXIV bis-2 et XXV-2), ont été calculées en un certain nombre de points intérieurs à l'ellipse

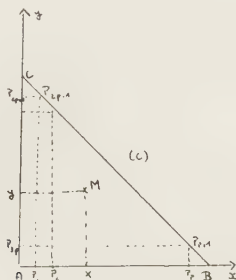


La comparaison des valeurs φ^* et φ^{**} (tableau 3) aux points proches du contour, c'est-à-dire numérotés 5, 10, 15, 20, 25 justifie la transformation faite sur l'intégrale. (surtout pour le point n° 10 ou 2 chiffres supplémentaires exacts sont obtenus).

(- voir tableau 3 page suivante-)

2°) Résolution du problème de Dirichlet - Cas d'un domaine triangulaire.

a) Remarques



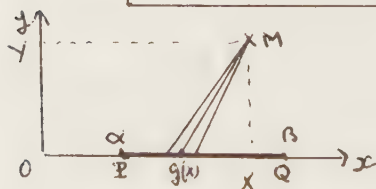
$$\begin{cases} AB = AC = d \\ BC = d\sqrt{2} \end{cases}$$

Point n°	φ_{exact}	φ_*	φ_{**}
0	1	1	1
1	0,98614323	0,98614320	0,98614321
2	0,94495695	0,94495696	0,94495694
3	0,87758256	0,87758257	0,87758256
4	0,78588726	0,78588720	0,78588722
5	0,67241224	0,67241307	0,67241234
6	1,0853448	1,0853448	1,0853448
7	1,1771315	1,1771315	1,1771315
8	1,2757663	1,2757663	1,2757663
9	1,3816721	1,3816720	1,3816720
10	1,4952886	1,4951499	1,4952704
11	1,0662817	1,0662817	1,0662817
12	1,1267457	1,1267457	1,1267457
13	1,1796527	1,1796527	1,1796527
14	1,2231181	1,2231165	1,2231174
15	1,2551173	1,2523886	1,2552309
16	0,92949071	0,92949069	0,92949071
17	0,85619385	0,85619386	0,85619387
18	0,78140012	0,78140008	0,78140011
19	0,70625374	0,70625384	0,70625377
20	0,63175662	0,63188776	0,63175435
21	0,92070821	0,92070819	0,92070821
22	0,84709781	0,84709784	0,84709784
23	0,77881440	0,77881439	0,77881437
24	0,71552058	0,71552058	0,71552059
25	0,65689576	0,65690159	0,65689623

Tableau 3

Le contour (C) présente ici trois points anguleux. La relation (XIX-2) qui donnait le potentiel en un point intérieur d'un domaine à contour régulier anguleux n'est plus ici applicable. (μ n'est pas définie aux points A, B et C). Nous considérons que le potentiel en un point M intérieur au triangle est la résultante des potentiels induits par les sources continuellement réparties, respectivement sur les segments AB, BC, CA

$$\Psi(M) = \varphi_{AB}(M) + \varphi_{BC}(M) + \varphi_{CA}(M)$$



Or nous savons que le potentiel induit, par un segment de droite PQ, sur lequel sont réparties uniformément des sources de densité ponctuelle $g(x)$, est en un point $M(X, Y)$.

$$\begin{aligned}
 \varphi_{PQ}(M) &= \frac{1}{\pi} \int_{\alpha}^{\beta} \frac{Y \, dx}{(X-x)^2 + Y^2} g(x) \quad \text{si } \begin{matrix} Y > 0 \\ \forall X \end{matrix} \\
 \varphi_{PQ}(M) &= g(X) \quad \text{si } \begin{matrix} X \in [\alpha, \beta] \\ Y = 0 \end{matrix} \\
 \varphi_{PQ}(M) &= 0 \quad \text{si } \begin{matrix} X < \alpha \text{ ou } Y > \beta \\ Y = 0 \end{matrix}
 \end{aligned} \tag{XXVI - 2}$$

b) Calcul de la densité $g(s)$

- Choix des points P_i -

Le choix des points P_i , auxquels seront attachées les densités ponctuelles g_i est imposé par la méthode de quadrature choisie. Pour cet exemple la méthode de Gauss (réf. [9]) a été adoptée. p points étant retenus sur chaque coté du triangle, le système à résoudre aura $N=3p$ inconnues g_i . Nous faisons croître, par convention, l'indice i dans le sens trigonométrique et notons

$$g(P_i) = g_i \quad \forall i = 1, \dots, N.$$

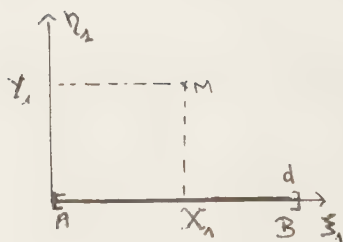
Et pour ne pas causer de confusions nous appelons $g^1(P)$ la densité relative aux points de AB, $g^2(P)$ celle relative aux points de BC, $g^3(P)$ celle des points de CA.

$$\text{avec donc } \left\{ \begin{array}{ll} g(P_i) = g^1(P_i) & 1 \leq i \leq p \\ g(P_i) = g^2(P_i) & p+1 \leq i \leq 2p \\ g(P_i) = g^3(P_i) & 2p+1 \leq i \leq 3p \end{array} \right.$$

$$\text{et } \left\{ \begin{array}{ll} g^1(B) = g^2(B) \\ g^2(C) = g^3(C) \\ g^3(A) = g^1(A) \end{array} \right. \quad \text{pour respecter la continuité de } g(P)$$

Pour calculer en M, les potentiels induits par les segments [AB], [BC] et [CA], nous attachons à chacun de ces segments un système d'axes de références $(\overrightarrow{A\xi_1}, \overrightarrow{A\eta_1})$, $(\overrightarrow{B\xi_2}, \overrightarrow{B\eta_2})$, $(\overrightarrow{C\xi_3}, \overrightarrow{C\eta_3})$, par rapport auquel M a pour coordonnées, respectivement (X_1, Y_1) , (X_2, Y_2) , (X_3, Y_3)

- Potentiel induit par AB



en appliquant (XXVI-2) on obtient

$$\varphi_{AB}(M) = \frac{1}{\pi} \int_0^d \frac{Y_1 d\xi_1}{(X_1 - \xi_1)^2 + Y_1^2} g^1(\xi_1)$$

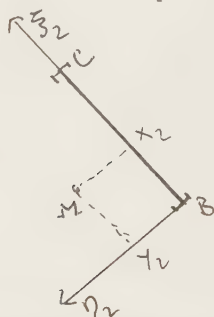
si $X_1 > 0$ et si $M \in (D)$

$Y_1 > 0$

$$\varphi_{AB}(M) = g^1(\xi_1) = g^1(X_1) \quad \text{si} \quad \begin{cases} X_1 \in [0, d] \\ Y_1 = 0 \end{cases}$$

si c.a.d. si $M \in [AB]$

- Potentiel induit par BC -

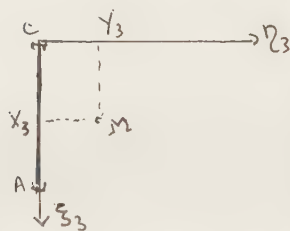


$$\varphi_{BC}(M) = \frac{1}{\pi} \int_0^{d\sqrt{2}} \frac{Y_2 g^2(\xi_2)}{(X_2 - \xi_2)^2 + Y_2^2} d\xi_2$$

si $M \in (D)$

$$\varphi_{BC}(M) = g^2(X_2) \quad \text{si} \quad M \in [BC]$$

- Potentiel induit par CA



$$\varphi_{CA}(M) = \frac{1}{\pi} \int_0^d \frac{Y_3 g^3(\xi_3)}{(X_3 - \xi_3)^2 + Y_3^2} d\xi_3$$

si $M \in (D)$

$$\varphi_{CA}(M) = g^3(X_3) \quad \text{si} \quad M \in [CA]$$

d'où le potentiel en M en remarquant que si M a pour coordonnées (x, y) relativement au système d'axe (OX, OY) (voir figure).

$$\begin{array}{lll} X_1 = x, & X_2 = (d-x+y) \frac{\sqrt{2}}{2}, & X_3 = d-y \\ Y_1 = y, & Y_2 = (d-x+y) \frac{\sqrt{2}}{2}, & Y_3 = x \end{array}$$

$$\begin{aligned}
\varphi_M(M) &= \varphi_{AB}(M) + \varphi_{BC}(M) + \varphi_{CA}(M) & (X_1, X_2, X_3, Y_1, Y_2, Y_3 \neq 0) \\
& & M \in (D) \\
\varphi(M) &= g^1(M) + \varphi_{BC}(M) + \varphi_{CA}(M) & M \in [AB] & (XXVIII-2) \\
\varphi(M) &= \varphi_{AB}(M) + g^2(M) + \varphi_{CA}(M) & M \in [BC] \\
\varphi(M) &= \varphi_{AB}(M) + \varphi_{BC}(M) + g^3(M) & M \in [CA]
\end{aligned}$$

- Construction de la matrice de résolution -

Cette matrice se construit en appliquant les trois dernières égalités de (XXVIII-2) aux points respectifs P_i , de AB, BC et CA, sachant qu'une intégrale $\int_a^b f(x) dx$ calculée par la méthode de Gauss est

$$I = \left(\frac{b-a}{2} \right) \sum_{i=1}^P H_i f\left(\frac{a+b}{2} + \frac{b-a}{2} t_i \right)$$

- les t_i et H_i , pour un p donné se trouvent dans des tables numériques (voir exemple réf. [9], appendice II pour $p \leq 17$).

Si on partitionne la matrice de résolution obtenue en 9 sous-matrices carrées de dimensions (p, p) , on constate que les sous-matrices diagonales sont des matrices identité.

Le calcul des valeurs propres extrêmes, nécessaires pour résoudre le système, a montré qu'elles étaient toutes positives et assez groupées, car

$$\forall n \quad \frac{\alpha}{\beta} < 10.$$

c) Calcul numérique du potentiel en un point intérieur de (D)

Le calcul approché du potentiel $\varphi(M)$ en $M \in (D)$ s'obtient en calculant numériquement la première relation de (XXVIII-2), le calcul des intégrales numériques s'effectuant à partir des mêmes points P_i , et en appliquant la même méthode de quadrature.

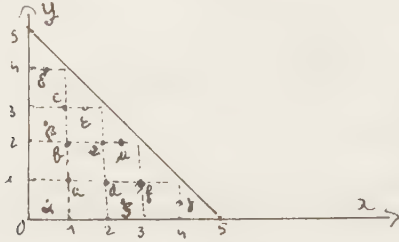
d) Résolution numérique - Résultats

Deux problèmes de Dirichlet ont ainsi été résolus numériquement sur le même domaine.

Le deuxième consistait en la recherche numérique des fonctions de Green $G(M, M')$ (par l'intermédiaire de la fonction $G_1(M, M')$ voir (II - 2.1.2.)) Ce problème avait le double intérêt d'expérimenter une méthode numérique de

recherche de telles fonctions et de la comparer avec une méthode expérimentale (recherche des fonctions de Green sur cuve rhéoelectrique). Le lecteur pourra comparer les résultats en se référant aux (réf. [17] et réf. [19]).

Le premier avait essentiellement pour but de vérifier la validité et l'intérêt de la méthode sur le plan numérique. Le potentiel exact était connu sous forme mathématique, d'ou comparaison possible entre solution exacte et solution approchée.



$$\left| \begin{array}{l} \varphi_{\text{exact}} = \cos x e^{-y} \\ d = 5. \end{array} \right.$$

Le tableau ci-dessous donne pour plusieurs résolutions, correspondant à un nombre total d'inconnues égal à 12, 27, 45 et 51, les résultats approchés obtenus - en des points intérieurs éloignés du contour

$$(\alpha, \beta, \gamma, \delta, \epsilon, \zeta)$$

- en des points proches de la frontière

$$(\alpha, \beta, \delta, \epsilon, \mu, \gamma, \xi)$$

Quelque soit la dimension de la matrice de résolution, les valeurs propres de la matrice restaient groupées sur le segment $[0.15, 2]$, $\implies \frac{\alpha}{\beta} \simeq 12$.

La résolution par méthode de Richardson s'est effectuée de façon très conforme aux prévisions théoriques, qui garantissaient 6 à 7 chiffres exacts pour la densité discrète obtenue.

Le tableau des résultats montre que pour ce domaine, il est nécessaire de traiter ce problème avec un nombre d'inconnues aux moins égal à 51, pour obtenir en tous les points une solution d'égale valeur. Une méthode de quadrature plus adaptée aurait sans doute été préférable. De plus il aurait fallu introduire plus de points P_i sur $[BC]$, qu'il ne l'a été fait. - (voir tableau page suivante) -

3°) Conclusion

Le dernier domaine étudiée est un domaine difficile sur le plan numérique du fait de la présence d'angles. Si les résultats obtenus sont moins précis que ceux du cercle, leur qualité acceptable montre l'intérêt que peut présenter l'application de la théorie du potentiel. Cet intérêt avait déjà été souligné (réf. [21]).

Le seul point délicat que présente cette méthode est la mise en équation du système. Le calcul de la densité $g(s)$, qui s'en suit, n'a jamais causé de difficultés, le rapport α/β , très faible, rendant la résolution du système linéaire par utilisation de la technique RAC, très rapide et très précise.

Point φ_{exact}		$\varphi_{\text{approché}}$			
		n = 12	n = 27	n = 45	n = 51
a	0,1987661	0,15704	0,199102	0,198768	0,1987660
b	0,07312197	0,0743546	0,0729812	0,0731234	0,731222
c	0,02690007	0,0545029	0,027747	0,0270922	0,026918
d	-0,1530919	-0,0909111	-0,151976	-0,153093	-0,153094
e	-0,05631935	-0,08020564	-0,0513453	-0,0558794	-0,0561211
f	-0,3641979	-0,32806	-0,36255	-0,363853	-0,364216
α	0,5322807	0,5732426	0,5295543	0,53218	0,532323
β	0,0720364		0,07355	0,072333	0,072170
δ	0,0160735	-0,0228	0,013138	0,0163045	0,016005
ε	0,003521798		0,0245886	-0,0028424	0,0015206
μ	-0,108423		-0,125167	-0,1181416	-0,11446
γ	-0,3964549	-0,357007	-0,4202213	-0,396034	-0,396815
ξ	-0,4859182		-0,5357571	-0,490414	-0,487947

Si le calcul du potentiel, à partir de cette répartition discrète de densités a parfois posé quelques problèmes, la technique de quadrature utilisée en était la cause.

BIBLIOGRAPHIE

[1] N. GASTINEL

Analyse numérique linéaire - Hermann - 1966

[2] G. FORSYTHE & W.R. WASOW

Finite difference methods for partial differential equations
John Wiley and Sons - 1960

[3] R.S. VARGA

Matrix iterative analysis - Prentice Hall - Englewood Cliffs -
New Jersey - 1962

[4] D. YOUNG

On Richardson's method for solving linear systems with
positive definite matrices - J. Math. Phys. - 32 - 243/255 -
1954

[5] E. STIEFEL

Relaxation methoden bester strategie zur lösung linearer
gleichungssysteme - Comm. Math. Helv. - 29 - 157/179 -
1955

[6] A. ANGOT

Compléments de mathématiques - Editions de la revue
d'optique - Paris - 1961

[7] W.L. FRANK

Solution of linear systems by Richardson's method -
J. Assoc. Comput. Mach. - 7 - 274/286 - 1960

[8] J. ALBRECHT

Iterations verfahren bester strategie zur lösung linearer
gleichungssysteme mit positiv definiter koeffizientmatrix
Z.A.M.M. (1963) 43 n° Specimen T4 - T8

[9] J. LEGRAS

Précis d'analyse numérique - Dunod - 1963

[10] J.F. MAITRE

Contribution à l'étude de la résolution numérique d'un
système linéaire par des stratégies gaussiennes - Thèse de
spécialité mathématiques - Besançon - 1964

- [11] Liste des matrices-tests - publication du laboratoire de
calcul - Besançon - 1966

- [12] J. LEGRAS
 Technique de résolution des équations aux dérivées par-
tielles - Dunod - 1956

- [13] J.M. VILLARD
 Etude de la déformation élasto-plastique d'une plaque
rectangulaire - Thèse d'ingénieur-docteur - à paraître 1968

- [14] Hideho NAGASAKA
 Error propagation in the solution of tridiagonal linear
equations - Information processing in Japan - Vol. 5 - 38/44
- 1965

- [15] J. COURBON
 Cours de résistance des matériaux - Dunod - 1954

- [16] LAURENT-GASTINEL-BOLLIET
 Un nouveau langage scientifique : ALGOL - Hermann - 1963

- [17] A. D'ANDREA
 Détermination expérimentale des fonctions de Green ; ap-
plication aux problèmes permanents - Thèse de doctorat de
spécialité mathématiques - Nancy - 1966

- [18] C. MONTET
 Etudes de techniques de discrétisation du problème de la
chaleur - Thèse de doctorat de spécialités mathématiques -
Nancy - 1966

- [19] Ph. NOEL
 Rapport des travaux exécutés sous contrat avec la D.R.M.E.
n° 387-64 - 1965-66

- [20] V.K. VLASOV & A.B. BAKUSHINSKII
 U.S.S.R. computational mathematics and mathematical
physics - Vol. 3 - Pergamon Press - 1963

- [21] V.N. FADDEEVA
 Computational methods of linear algebra - Dover Publication
- New-York - 1959

TABLE DES MATIERES

	Pages
INTRODUCTION	1
1ère Partie : METHODES ET TECHNIQUES	
1.1. - Méthodes	
1.1.1. Position du problème	3
1.1.2. Notations utilisées	3
1.1.3. Détermination des polynômes matriciels $[P_p(A)]$ et $[Q_{p-1}(A)]$	4
1.1.4. Estimation de l'erreur commise	6
1.1.5. Choix du polynôme de convergence $[P_p(A)]$	8
1.1.6. Résolution d'un système linéaire dans le cas général	16
1.2. - Techniques	
1.2.1. Introduction	21
1.2.2. Technique R.A.C. : utilisation des racines du polynôme $P_p(x)$. Comparaison avec la méthode de Richardson	22
1.2.3. Technique R.E.C. : utilisation de la relation de récurrence liant trois polynômes consécutifs de Tchebycheff Comparaison avec la méthode de Stiefel-Frank	30
1.3. - Problèmes de l'optimisation	
1.3.1. Position du problème	41
1.3.2. Détermination du degré P du polynôme $[P_p(A)]$	42
1.3.3. Influence des erreurs d'arrondi sur un procédé itératif	47
1.3.4. Contrôle des erreurs d'arrondi	52
1.3.5. Résolution d'un système linéaire "franc"	53
1.3.6. Résolution d'un système linéaire "non franc"	56
1.4. - Résolution de système test	
1.4.1. Matrice tridiagonale	56
1.4.2. Matrice pleine n° 1	60
1.4.3. Matrice pleine n° 2	63
1.4.4. Conclusion	68

1.5. - Compléments : Détermination des valeurs propres extrêmes. Utilisation des polynômes de Tchebycheff pour la recherche de la valeur propre la plus petite en module	
1.5.1. Nécessité d'une bonne connaissance des valeurs propres extrêmes	69
1.5.2. Détermination numérique de α , valeur propre la plus grande en module, par méthode itérative	69
1.5.3. Détermination de β , valeur propre la plus petite en module de $[A]$	71
1.5.4. Application numérique	79

2ème Partie : *QUELQUES APPLICATIONS AUX EQUATIONS AUX DERIVEES PARTIELLES*

II-1 - Résolution d'un problème d'élasticité thermique	83
II-1.1. Problème de la lingotière : Mise en équation	83
II-1.2. Problème de la lingotière : Résolution numérique	89
II-1.3. Résultats obtenus	96
II-2 - Méthodes numériques de résolution du problème de Dirichlet	105
II-2.1. Superposition finie de solutions harmoniques	105
II-2.2. Utilisation de la méthode du potentiel	113

BIBLIOGRAPHIE	129
---------------	-----

- DEUXIEME THESE -

Propositions données par la Faculté

TRAITEMENT DES LISTES EN PL/1

Vu et Approuvé,
NANCY, le 26 janvier 1967

Le Doyen de la Faculté
des Sciences de Nancy,

J. AUBRY

Vu et Permis d'imprimer
NANCY, le 30 janvier 1967

Le Recteur,
Président du Conseil de
l'Université,

P. IMBS

Cet exemplaire est imprimé
à des fins particulières et ne
peut en aucun cas être vendu.

Achévé d'imprimer le 20 Octobre 1967
à l'atelier de Documentation Universitaire
3, Rue Sellier, 3 NANCY

Dépot légal N° 302 Faculté des Sciences Editeur
4ème trimestre 1967

